manager can delegate to an outside consultant and merely await the report—although an outsider can facilitate the discovery of answers and the reduction of internal barriers to their expression. Collectively, participants must have a wide range of knowledge, of interest, and of technical expertise—more than is likely to be found in any one person. The best procedure for organizational need analysis may be to form a task force of bright people who know the organization from a variety of perspectives, augment them as necessary with hired specialists in various problem solutions or in discussion processes, and let them study, question, argue, and arrive at their best collective judgments.

JOB ANALYSIS

When organizational needs require improved personnel decisions for people on specific positions, jobs, or groups of jobs, job analysis (or position analysis) is necessary. Jobs are analyzed to understand them clearly enough to know which variables or performance constructs should be predicted and to identify variables or constructs that might be effective predictors—that is, to develop predictive hypotheses.

Some Definitions

Following McCormick (1979), with some additions and liberties of my own, here are some more or less standard definitions:

Position: The duties and tasks carried out by one person. A position may exist even where no incumbent fills it; it may be an open position. There are at least as many positions in an organization as there are people.

Job: A group of positions with the same major duties or tasks; if the positions are not identical, the similarity is great enough to justify grouping them. A job is a set of tasks within a single organization or organizational unit. This definition does not preclude flexibility. Members of a self-contained work unit may, on any given day, be doing different tasks, but each member may also be expected to do on another day any task the group as a whole must do.

Occupation: An occupation is a class of roughly similar jobs, found in many organizations and even in different industries. Examples include attorney, computer programmer, mechanic, and gardener.

Job family: A group of jobs similar in specifiable ways, such as patterns of purposes, behaviors, or worker attributes. Pearlman (1980) applied

the *family* concept to occupations, but the term is usually applied to sets of jobs within an organization.

Job analysis: Job analysis is a study of what a jobholder does on the job, what must be known in order to do it, what resources are used in doing it, and perhaps the conditions under which it is done. What the jobholder does may be defined in several ways: as tasks, classes of duties or responsibilities, broad activities, or general patterns of behavior. What must be known includes job knowledge and job skills. What resources are used may include those the person may bring to the job (relevant experiences, general abilities, or other personal characteristics), tools and materials used (e.g., manuals or handbooks, supplies, or equipment) or the work products of other jobs or work units.

Element: The smallest feasible part of an activity or broader category of behavior or work done. It might be an elemental motion, a part of a task, or a broader behavioral category; there is little consistency in meanings of this term.

Task: A step or component in the performance of a duty or activity. A task has a clear beginning and ending; it can usually be described with a brief statement consisting of an action verb and a further phrase.

Activity (or *responsibility* or *duty*): A relatively large part of the work done in a position or job. It consists of several tasks related in time, sequence, outcome, or objective. A clerical example might be "sorting correspondence" or "handling cash" or "preparing reports." All tasks grouped under these activities are done for a common end. One task in correspondence sorting might be "identify letters requiring immediate response." Putting together a report includes such tasks as laying out or formatting tables and charts, typing text, typing tables or charts, proofing for errors, and perhaps duplicating, collating, and binding copies of the reports. Activities and tasks are both components of jobs, but activities are usually considered more general, more encompassing.

Essential function: A term introduced in the *Americans With Diasabilities Act (ADA)*, which defines a "qualified individual with a disability" in part as one who "can perform the essential functions of the employment position that such an individual holds or desires" (Schneid, 1992, p. 28). The meaning of many terms in the ADA, including this one, waits on court decisions and developing case law. In the meantime, EEOC regulations identify three considerations: (a) whether the position exists for the purpose of carrying out the function, (b) whether the number of employees who can perform the function is limited, and (c) whether the function is highly specialized so that people are

hired because of their special expertise or ability to carry out the function (Schneid, 1992, pp. 33–34).

Job description: A written report of the results of job analysis. A job description is usually narrative, sometimes given in a brief summarizing paragraph. It may be more detailed. Where job analysis was done by survey methods, the description may include listings of task statements found to define or characterize the job being studied, along with statistical data.[3]

Job specification: Required qualifications for the job (or position), as revealed in the job description. Depending on the job or job category, specifications can include legal requirements (age, licenses, residency, etc.), education, skills, or perhaps assessment standards (although the latter requires research beyond the job description).

Detail Versus Generality in Job Analysis

In job analysis, a job as a whole is analyzed into component parts; the level of detail can vary widely. Detailed statements may be best for developing training programs, but more general statements are more useful for identifying criteria and predictors for selection (Lawshe, 1987).

Clarity counts more than detail. Lawyers and courts want more detail than is useful. Too much detail can muddle matters; what is needed is a clear enough understanding of the job to move on to the next step, the development of one or more predictive hypotheses. These require the wisdom, insight, and even introspection of people who know and understand the job. Job analysis can tap the wisdom and knowledge of job experts. Highly detailed, cover-all-bases, formal job analysis may not be needed at all—except possibly for convincing others that the analysis was done well.

Information *needed* is not necessarily the information *desired*. In an age of litigation, actions are governed as much by what is prudently filed away as by what is actually needed. Fine details may not be needed for any purpose beyond a trial. Failure to convince a trial judge that the job analysis was "adequate"—lots of questions asked, results recorded in a lengthy job description along with lots of statistical analyses—may be the

[3]In a peculiar pair of definitions, the *Uniform Guidelines* defines job analysis as "a *detailed* statement of work behaviors. . . ." and job description as "a *general* statement of job duties. . . ." (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978, p. 38307, italics added). Obviously, these definitions were written without much regard for the meanings of the verbs *to analyze* or *to describe*. In my judgment, the failure to recognize a difference between the process of analyzing something and the description of the results has resulted in mischief in court cases.

11

Assessment by Testing

Of all assessment methods, testing has the best foundation in research, measurement theory, and the development of standards of evaluation. Other methods of assessment may be preferred for some purposes, but the development and use of tests provides a prototype for the development, use, and evaluation of assessment in other forms.

A *test* is an objective and standardized procedure for measuring a psychological construct using a sample of behavior. A test is objective in that responses can be evaluated against external standards of truth or of quality—correct or incorrect, or better or poorer than a standard. Measuring implies quantification. Tests are scored quantitatively, with measurable precision, on numerical scales representing levels of a construct to be inferred from the scores. A *construct*, as I use the term, is a fairly well-developed idea of a trait; most constructs in testing are abilities, skills, or areas of knowledge. Tests use a standardized procedure with the same stimulus component for all test takers. *Standardization* refers primarily to controlling the conditions and procedures of test administration, that is, keeping them constant—unvarying. If scores from different people are to be comparable, they must be obtained under comparable circumstances. If people tested in one room have 30 minutes in which to complete a test, and those in another have only 20 minutes, neither the circumstances nor the scores are comparable. Any circumstances of test administration potentially influencing scores should be standardized. More than anything else, it is attention to standard procedure that distinguishes testing from other forms of assessment. The distinction is fuzzy. In this chapter, I describe a variety of procedures for assessing KSAs,

ranging from highly standardized tests to assessments with little or no standardization, with no clear line distinguishing tests from other assessments procedures.

Defining a test as a sample of behavior means that the examinee is not passive but does something. In other kinds of testing (e.g., blood tests) the object of measurement sits passively while something is done to it. In psychological tests, the examinee responds to test stimuli by writing answers to questions, choosing among options, recognizing or matching stimuli, performing tasks, ordering objects or ideas, or producing ideas to fit requirements—and this is not an exhaustive list.

TRADITIONAL COGNITIVE TESTS

Cognitive tests allow a person to show what he or she knows, perceives, remembers, understands, or can work with mentally. They include problem identification, problem-solving tasks, perceptual (not sensory) skills, the development or evaluation of ideas, and remembering what one has learned through general experience or specific training. They include intelligence tests, achievement tests, and job knowledge tests, among others.

Some History: Oral Trade Tests

Oral trade tests were among the earliest of employment tests. Many were developed in the Army during World War I and in industrial locations in the years immediately after (Chapman, 1921; Link, 1919; Poffenberger, 1927). Trade tests were needed, in part because many applicants genuinely thought they had expertise in a trade when, in fact, they knew only a limited aspect of it; oral tests were needed because many applicants could not read. The problems increased in the depression years and were addressed with excellent test development work in the United States Employment Service (USES). The work, described by Osborne (1940), is unfortunately no longer well-known.

USES job analysts in different regions made detailed observations and developed questions about a wide range of topics fundamental to the trade. Questions and the correct answers were reviewed regionally and nationally for adherence to specified principles, such as brevity. Research compared scores in three groups of subjects. People in the *A* group were highly skilled in the trade with at least four years of post-training experience. The *B* group included apprentices, helpers, or beginners. The *C* group consisted of people in other occupations whose work gave them contact with workers in the targeted trade. The proportion of correct

FORM I

Score	Expert Bricklayers (n = 65)	Apprentices and Helpers (n = 25)	Related Workers (n = 35)
15	xx		
14	xxxxxxxx		
13	xxxxxxxxxxxxxxxx		
12	xxxxxxxxxxxxxxxxxxxx *		
11	xxxxxxxx	x	
10	xxxxx	xx	
9	xx		x
8	xx	x	
7		xx	
6		x	x
5		xxxxx *	
4		xxxxx	x
3		xx	xxx
2		xx	xx
1		xxx	xxxxxxxxxxx *
0			xxxxxxxxxxxxx

FORM II

Score	Expert Bricklayers (n = 65)	Apprentices and Helpers (n = 25)	Related Workers (n = 35)
15	x		
14	xxx		
13	xxxxxxx		
12	xxxxxxx		
11	xxxxxxxxxx		
10	xxxxxxxxxxx *		
9	xxxxxxxxxxx		
8	xxxxx	x	x
7	xxxx		x
6	x	x	x
5		xx	
4	x	xxx	x
3		xxxxx	xxx
2	x	xxxxx *	xxxxx
1		xxxx	xxxxxxxxxx *
0		xxxx	xxxxxxxxxx

* median score

FIG. 11.1. Distribution of scores on two forms of the oral trade test for bricklayers. From Osborne (1940).

answers should be significantly higher in group *A* than in groups B or C.[1] An example of the results for bricklayers is shown in Fig. 11.1.

Traditional Tests

Most tests now used are called paper-and-pencil tests, but materials do not define traditional tests. The defining features of traditional tests are that they are well-standardized, that their items can be reliably scored, and that they can be administered to groups of people.

[1]Later, as experience demonstrated that response patterns varied little between groups *B* and *C*, only one "control group" was used. It was designated group *C*, mainly apprentices, helpers, or beginners but augmented as needed by workers in related occupations.

Commercially Available Versus Homemade Tests. It is almost always cheaper to buy a test than to develop one; moreover, commercial publishers are likely to do a better job of writing, calibrating, and evaluating items and empirically evaluating the resulting test. When there is a choice, commercially available tests have clear advantages.

A commercial test may have less face validity, and therefore less acceptance by the people tested, than a locally developed test that refers explicitly to specific jobs or sets of jobs within the organization. Job-specific local tests developed by people well-trained in psychometrics can be as reliable and valid as commercially available ones. One study paired three subtests of the *Differential Aptitude Test Battery* (DAT) with related job specific tests (Hattrup, Schmitt, & Landis, 1992). For example, the DAT Verbal Reasoning test, a measure of the verbal comprehension factor, was paired with a technical reading test based on manuals used on the job. Confirmatory factor analysis showed that the same constructs were measured in each of the three pairs of tests. Hattrup et al. (1992) concluded that test users do not gain much, psychometrically, by building homemade, job-specific tests, even good ones, but that they do not lose anything, either, and may gain considerably in testing program acceptance. I see a third implication. No matter how much a test developer tries to make particular tests highly specific to particular uses, general cognitive constructs still account for most of the variance. Those who think they are doing things that are new or highly specific may only be fooling themselves.

Tests may have to be developed by local people to serve local purposes. I am not as skeptical of homemade tests as I was before studying a test of electronics knowledge developed by an inspection supervisor. The company, fearful of litigation but unwilling to challenge the supervisor, called for an outside evaluation of the test. It met every reasonable expectation. I have seen other examples of psychometrically good homemade tests (and some that were not). Well-informed job experts can make good tests, especially if helped by someone trained in test development principles.

Traditional Item Types. Responses to questions on the earliest tests, such as oral trade tests, were the examinee's own, not chosen from a limited set. These free responses are not usually called traditional; in general, *traditional* items permit reliable scoring. It is reliable scoring that typifies tests called traditional, not the response format. Figure 11.2 gives examples of reliably scored item types.

Multiple-choice items are prototypical traditional items; they provide reliable, valid tests. They allow an examinee to choose one correct (or best) response from perhaps three to six options. They are versatile; they can test

Sentence Completion

General assumptions about the sources of random error variance in a set of test scores are involved in the estimation of _____.

Short Answer

Estimates of reliability are intended to partition total test score variance into two components, What are they?

True-False

T F Reliability refers to a contaminating source of variance in a set of test scores.

Multiple Choice

An estimate of reliability that treats the choice of items presented as a source of error variance is known as

A. a coefficient of stability
B. a coefficient of equivalence
C. an internal consistency coefficient
D. a conspect reliability coefficient

Multiple Choice

Several assessment methods are used in a day-long assessment process: a 50-item multiple-choice test of job knowledge, a performance exercise rated by observers, a brief essay test with 5 items, and one large essay. Standard scores on the parts are added to provide a composite overall assessment. Considering both reliability theory and experience with estimation methods, which of the following estimates of reliability do you consider most appropriate for the overall score?

A. Internal consistency within the total assessment procedure
B. Stability of performance from time to time during the day
C. Equivalence of the components used in developing the overall score
D. Level of agreement of observers and scorers

FIG. 11.2. Examples of objective item types (i.e., types of items that incur little or no random error in scoring) for an examination on measurement theory.

for grasp of factual information (at abstract or at simple levels) or for abilities to reason from given premises, calculate, evaluate optional courses of action, identify causes or effects or associations, detect errors, infer operating principles, or comprehend principles, sequences, or arguments.

The multiple-choice format has many advantages. It permits testing at a variety of levels of cognitive functioning, from simple recognition to the analysis of problems and evaluations of solutions. It permits wider sampling of relevant content than possible with free responses, thereby providing better coverage of content domains. It may encourage guessing, but it reduces the bluffing encouraged with some constructed response forms. It reduces subjectivity in scoring and generally has higher reliability.

Multiple-choice testing is often criticized, usually as superficial. Superficiality is the item writer's fault; it is not inherent in the format. Figure 11.2 shows two multiple-choice items. One demands only recall; the other

demands understanding of theory and accumulated results of applying the theory. Many criticisms, such as that multiple-choice formats inhibit the expression of creativity, can be valid—if the purpose of testing is to assess creativity. It may be. It is unwise in selecting managers, for example, to select those unlikely to show any originality (although it happens); the assessment of potential managers should include assessments of abilities to think unconventionally, to produce ideas readily and in volume, and to change ways of looking at problems—all part of creativity. However, managers should also be able to do arithmetic in their heads (to spot substantially wrong computations), choose words to convey special meanings, or perceive details quickly and accurately—all of which may be assessed best with traditional multiple choice tests. Choosing a method of assessment should be based on its purpose, not on some overgeneralized preference for one sort of test over another. Other item types in Fig. 11.2 have less to recommend them and are subject to similar criticisms. Perhaps that is one reason why the multiple-choice items are more widely used.

PRINCIPLES OF TEST DEVELOPMENT

People responsible for assessments in organizations need to understand how tests are developed. This includes people who must pass judgment on a proposed testing program, administer such programs, take administrative actions relative to or based on such programs, and develop ways to assess aptitude, performance, knowledge, or ability of others. Many assessment procedures are developed locally, even if not tests, and they will be better assessments if developed with awareness of test development principles.

Testing implies assessment with reliably fine gradations along a scale. Not all assessment programs seek that level of measurement precision, but at any level, an assessment procedure important to the organization deserves to be developed with care and understanding—even if it seeks no more than assignment of assessees to a few ordinal categories. The process of developing traditional tests illustrates basic principles applicable as well to other assessment methods.

The Basic Construct and Content Domain

Conceptual Definition of Purpose of Measurement. Test development starts by clearly saying what is to be measured, the construct that gives intended meaning to the scores. When the intended construct is vague, when one has no clear idea of what is to be measured, one cannot know whether it has been measured well. The idea of a construct need not be

daunting. For practical purposes, a construct is any idea or concept of an attribute of people, jobs, behavior, environments, or other entities. Clarification of the idea may distinguish it from some other ideas and relate it to still others. A clear idea is more than a name for the construct; it is an idea defined in detail. Its definition should clearly identify its boundaries, that is, what the construct is and what it is not, and there should be some unity of concept within those boundaries. When fully defined, with boundaries and distinctions, it becomes a theory of the attribute to be measured—essential for understanding the basis for practical decisions. It should be potentially useful, from either an organizational or a scientific perspective—preferably both. To be useful, it must imply important individual differences, be subject to empirical quantification, and remain reasonably stable over a substantial time period.

Test Specifications. Construct boundaries enclose a *universe of admissible observations* (Cronbach, Gleser, Nanda, & Rajaratnam, 1972, p. 20). Boiled down, that phrase means that a test developer specifies some observations that fit the construct and some conditions or circumstances appropriate for making them. As the aphorism about skinning cats has it, a construct can be measured by more than one kind of observation or circumstance. The words *universe* and *domain* are often used interchangeably, but I find it useful to distinguish between them, considering a domain a nonrandom sample of a defined universe (Guion, 1979; Lawshe, 1975). Knowing construct boundaries and test purposes can aid in defining a universe of all acceptable kinds and conditions of measurement that may satisfy those purposes; specifying a test content domain narrows the universe and provides a test plan or set of specifications. The following should be part of test planning that continues to clarify the construct:

1. *Specify the kinds of behavior to be observed and the kinds of stimulus materials that will elicit that behavior*—in short, specify test content. Test content is not simply information or tasks; it includes all stimulus characteristics (form, time limits, and other aspects of standardization). Some tasks or behavior fit the construct better than others, but good choices usually require long thought. Choices might be based on psychometric experience (knowing tasks that have a good track record for the construct), practical considerations (e.g., cost, time requirements, or likely attitudes of the examinees), or expected social or legal consequences.

2. *Specify intended inferences as norm- or domain-referenced.* Most test scores are interpreted normatively—in standard deviations above or below the mean, or in centile units—in the distribution of scores from a specific group. Such interpretations answer the question, "How well did I do relative to other people?" An alternative question, "How well did I

do relative to some standard of excellence?" is answered in domain-referenced interpretations. Domain-referenced inferences should be used for personnel decisions more often than they are. For norm-referenced testing, item statistics are particularly important. If domain-referenced interpretations are intended, domain sampling rules are more important.

3. *Specify test components and their intended psychometric characteristics.* Test components are usually items or sets of items (testlets).[2] Characteristics such as difficulty or discriminability can be targeted in advance. For normative interpretations, psychometric theory suggests an average difficulty of .5, but that is not necessarily the best level for every application. If constructed responses are used, or if the test is to differentiate only among the highest scoring candidates, perhaps the average difficulty level should be .3 with a range from .1–.5.

4. *Specify the medium of presentation.* Options include presenting stimlus components orally, on paper, on audio- or videotape, or via a computer. The choice should be made after carefully considering alternatives, and it should make sense in the light of the construct measured.

5. *Specify the medium of response.* There is no compelling reason for the medium of presentation to be the medium of response, although (for the convenience of almost everyone, including examinees) that is typical. Written questions can be answered orally; responses to videotaped situations can be entered on machine-scored answer sheets. Whatever, the response should not violate the nature of the construct.

6. *Specify constraints on responses.* Should responses be constrained or relatively free? The answer should depend on the theory of the attribute, not on the convenience of testers or managers.

7. *Specify appropriate population characteristics.* Some reasons are obvious. Verbal items should be written to be understood by people in the intended population. A test item for tool and die makers may ask about the properties of metals—but not phrased in the jargon of metallurgical engineers. A less obvious but more important reason is that pilot studies should use relevant samples.

8. *Specify content allocations.* Domain boundaries may include several components, some more important, complex, or informative than others. Boundary judgments imply a desired distribution in the final test and the proportion of items, testing time, or points given to each. The intended allocation may be hard to keep during development; items for some topics may not have the desired difficulty statistics, or judgments of the content

[2]Testlets were introduced as homogenous item sets used in computerized adaptive testing (Wainer & Kiely, 1987); I use the term more broadly, meaning any internally consistent subset of items within the larger set that is the test.

relevance for some topics may be less consistent than planned; whole component topics can be lost if this specification is not met.

9. *Specify time limits, if any.* Practicalities may impose some time limits; a test given during a training class, for example, may have to be completed, with instructions and collection of papers, within a 50-minute period. Time limits may be set because speed is part of the construct. *Speeded* tests differ from *power* tests, the latter show what examinees can do without the constraints of imposed time limits. If the intended construct is defined by power, but administrative considerations impose a time limit, the specified time should allow nearly all examinees to complete the test; this effectively limits test length.

10. *Specify other standard circumstances of testing.* Establishing time limits, specifying instructions, offering sample items, limiting explanation, arranging the items in sequence, sizing the type, adjusting the video display resolution or color-coding materials for assembly, establishing scoring procedures, and the required qualifications for test administrators, among others, may need standardization. Where specifications require pilot studies, the test development plan should include plans for them.

These 10 points are not exhaustive, but they offer a flavor of things to consider in test development—the statements, decisions, and clarifications needed for a workable plan. A test plan, like a house plan, can be changed as the work moves along; a particular part of the plan might prove ambiguous, esthetically unpleasing, too time consuming, or have some other unanticipated flaw. For the most part, however, both houses and tests are constructed according to the basic plan, even though deviations in details can be anticipated. However, the plan should include procedures for recording deviations and the reasoning behind them. Such records are useful in evaluating psychometric validity.

Developing Items or Other Components

Good professional judgment is required for developing any kind of item, and good judgment requires experience. Because test development experience is greatest for multiple-choice items, I concentrate on them.

A multiple choice item has three parts: stem, a correct response, and a set of distractors. One way to write a multiple choice item is to (a) write a true statement, (b) delete a word or phrase as one would in writing a completion item, (c) write some words or phrases that would be unacceptable but plausible answers in a completion item, and (d) list them, with the correct one, as a set of options. These items assume that people who know the correct answer will choose it and that there is no widespread *mis*information on the item's topic (Horst, 1966). The first assump-

tion is fairly safe in most employment settings. The second assumption is less safe, especially with some job knowledge items; not often, but often enough to merit concern, people will talk about experiences, build on them, and build on the comments of others so much that they know things that simply are not true. Items influenced by widespread misinformation can generally be identified in item analysis.

Several authors have offered useful rules for writing multiple-choice items; they should be studied carefully before trying to build a multiple-choice test (Ebel, 1972; Hopkins & Stanley, 1981; Millman & Greene, 1989; Osterlind, 1989; Thorndike & Hagen, 1955). Examples include:

1. Give each item some "face validity." Make items obviously relevant to the announced purpose of the test; use language that is appropriate in word choices and reading levels.
2. Be sure the item content is suitable for the purpose of the test and for the examinee population intended.
3. Write in clear and simple language and style; keep vocabulary level as simple as the problem allows.
4. Avoid negative words or words that exclude something (e.g., not, except); if they cannot be avoided, emphasize them with capital letters *and* italics or boldface type.
5. Be sure that there is just one correct (or best) answer; be sure that options are false (or, if all are partly true, that there is a clear principle for declaring one better than the others). If the item involves controversial matters, ask for the position held by a specified authority (or, better, ask about the nature of the controversy).
6. Be sure the problem for the examinee is clear in the stem. Phrasing the stem first as a question may help the test developer clarify the examinee's task, and that clarity may spill over to an item edited to another form such as an incomplete sentence.
7. Put as much of the item as possible in the stem, avoiding repeated use of a phrase in each response option (unless repetition provides more clarity); options should be as brief as possible. The stem should be as brief as consistent with clarity; excess verbiage creates ambiguity.
8. Avoid *specific determiners*—cues within items giving away the correct answer, that help an examinee who lacks the desired knowledge identify the preferred option anyway. An example is a stem ending with the article "an." If only one option begins with a vowel, that option is most likely to be chosen. Other specific determiners include variations in length or grammatical structure of the options, use of the same word in the stem and an option, implausible conditions such as never or always, or consistent placement of correct options within the sets.

9. Keep the number of options constant. If this cannot be done, vary the number of options early in the test to avoid establishing spatial sets.

10. Be sure distractors are plausible. In a good item, incorrect responses are distributed rather evenly. Good distractors can come from a pilot study giving stems as questions and calling for written responses.

11. Keep distractors similar in content. If all options refer to different aspects of the problem, each one is a true–false item; answers depend more on cleverness in eliminating options than on knowledge or understanding of the test material.

12. Keep options independent of each other; if two options are merely different ways to say the same thing, they can both be discarded quickly by an alert examinee.

13. Avoid "none of the above" or "all of the above" as options when examinees are to choose the best rather than the correct answer.

14. Keep items independent; do not let information in one item provide a cue to the answer to another one.

Some psychometrics text books give similar rules for other item types. For some, however, there has been too little experience to form generally acknowledged rules. When a relatively rare item form is needed, guidelines or rules developed in advance can make item development more systematic, even if the wisdom of individual rules is uncertain.

Pilot Studies

Choices are made at nearly every step in test development. Some can be made rationally, considering one's options and the relevant arguments for or against them. Other choices need data, and sometimes data must be developed locally. Choices need to be informed by answers to one or both of two kinds of questions: Is this working? What if?

Preliminary Studies. Pilot studies need not be elaborate or sophisticated; they may be simple trials of procedural ideas or choices. Simple studies can usually show whether enough time is being allowed to complete a power test, or whether instructions are clear enough. Structural glitches may be suspected and lead to "is it working" questions. The more novel the test, the more important such questions are. Some of them may require full-scale experiments, but many of them can be answered by trying out the instructions, or the time limits, or purely physical aspects of the test to see if they work. A useful study can be done asking a few people to think out loud as they take the test, and the listening test developer can learn where instructions go awry, or how

distractors do not distract. The trials should, of course, use appropriate samples, but the samples need not be large. Such trials are easy, but should not be dismissed as trivial luxuries. It is terribly arrogant and even self-defeating, to assume that one's expertise assures a good, workable plan.

Conventional Item Analyses. Likewise, it is terribly arrogant and surely self-defeating to assume that items developed or chosen work as intended; item analysis is neccessary. Where possible, samples should be large enough to provide reliable item statistics, they should be similar to the population for whom the test is intended, and the trait measured should be distributed somewhat as it is in that population. These are tall orders, unlikely to be fully satisfied in most do-it-yourself test development, but they provide goals to be approximated as well as reality permits. Doing the best one can, even if imperfectly, is preferable to doing nothing.

Two kinds of item statistics are traditionally computed. The easier of these is item difficulty, computed (in reverse) as the proportion of those tested who give the keyed response. Conventionally, .50 is considered an optimal proportion, but other target levels can sometimes be better. Conventional wisdom also suggests that item difficulties should be essentially equal and that item responses should be highly correlated (to produce an internally consistent test).

Such "wisdom" needs to be tempered with good judgment. If all items are perfectly correlated, and if all items have difficulty indices of .50, then this test yields a 2-point distribution where half of the scores are zero and the other half are perfect. The result is classification, not measurement. To develop a fairly internally consistent, homogeneous test with scores distributed along a scale, item difficulties must vary. If the intended examinees define a general population, perhaps the average around which that variation occurs should be .50. For homemade tests, however, the intended use may require that distinctions be most reliable at higher or lower levels of the trait, and the average item difficulty around which specific item difficulties are distributed may therefore be higher or lower.

The other traditional item statistic is a discrimination index. For dichotomously scored items, a preferred index is a point biserial correlation coefficient. The criterion in computing this coefficient is usually the total test score (minus the item being analyzed); sometimes an external criterion is used. Both were used in developing the *Purdue Mechanical Adaptability Test*. Internal consistency was one goal, so one item analysis pitted item responses against total scores. Another goal was to have a test that measured mechanical knowledge, not general intelligence; therefore, a second item analysis pitted item responses against scores on a test of general mental ability. To be retained, an item had to correlate well with

total score but not at all well with general mental ability. I do not know why this excellent example is so rarely followed.

Ideally, item statistics are computed in two or more independent samples; only those items in the final test that meet the specifications in double cross validation are retained. However, one sample large enough for reliable statistics is hard to get, let alone two. A less sophisticated approach divides one sample into both a high criterion group and a low criterion group. Differences in proportions giving the keyed answers in these groups is a simple item discrimination index.

Item Analysis by Item Response Theory. An IRT model with two or more parameters for the item characteristic curve provides corresponding item statistics. The slope of the curve at the point of inflection provides the item discrimination index (*a*) and the location of that point along the Θ scale is the item difficulty index (*b*).

Reliability and Validity Analyses. Further pilot studies provide data for the preliminary evaluation of tests. Some of these might be major undertakings; for example, a full scale generalizability analysis encompassing the traditional aspects of reliability (stability, equivalence, internal consistency, and interscorer reliability) requires large samples and careful planning. Full scale construct validation requires reliable testing of several confirmatory, disconfirmatory, and competing hypotheses. Even where large scale studies are not feasible, however, they can be outlined conceptually and doing so can suggest plausible sources of random or contaminating error. Even modest research can permit opportunity to address serious problems while there is opportunity to change the test before making it operational.

APPROACHES TO TAKING TESTS

Individual differences in ways people take tests may be a source of irrelevant variance. Differences in test-taking strategies have long been recognized, at least anecdotally, but data are sparse. The relevant literature seems limited to attitudes toward testing, the influence of cognitive styles, and the concept of test-wiseness.

Attitudes

A 9-scale *Test Attitude Scale* (TAS), developed and used by Arvey, Strickland, Drauden, and Martin (1990), accounted significantly for variance on other tests; mean attitudes differed significantly for job applicants (as

in predictive or follow-up validation designs) and incumbents (as in concurrent designs) on seven of the scales. Scores on some scales correlated significantly (but not highly) with race, sex, or age.

Schmit and Ryan (1992) used a TAS composite score (reflecting positive test-taking dispositions) in a study using ability and personality tests to predict GPAs of college students. TAS scores moderated predictions—but in different directions. A positive motivation composite score was associated with higher validity for the ability test; a negative motivation composite was associated with higher personality test validity.

Test-Taking Styles and Strategies

Other research has studied individual differences in test-taking strategies such as changing answers. Contrary to popular thinking, people who change answers usually improve their scores, especially if they had initially high scores (Pike, 1978, p. 31). Some people answer all multiple choice items in sequence, but most skip some items and return to them later. In one study, more than 70% of the students omitted items for later consideration (Schwarz, McMorris, & DeMers, 1991).

Willingness to guess may account for some variance in scores, perhaps associated generally with individual differences in risk-taking propensities (Ben-Shakhar & Sinai, 1991; Pike, 1978). Individual differences in cognitive style also contribute some systematic error variance. Acquiescent response styles (tendency to say "yes" or "true" regardless of wording) pose a problem for true–false items. Positional response styles pose a similar problem in multiple-choice items. These are tendencies to choose either extreme or nonextreme responses: in a four-option item, some people choose extreme *A* or *D* positions when guessing; others choose the central *B*–*C* positions. Other cognitive styles also influence scores. Long ago, French (1965) identified several problem-solving styles that could be subsumed under the contrasting poles of analyzing versus a more global way of perceiving. More recently, Armstrong (1993) found field dependence versus field independence related to performance on poorly constructed test items, especially those with specific determiners. Field dependent people must be taught to look for cues in items, and to know that they are not part of the "total field" of the item that field independent people see spontaneously.

People may also differ in their sensitivity to language and language patterns. Items containing specific determiners are indeed poorly written—but not equally so. The most blatant of them seem likely to point to the correct answer for everyone. More subtle ones, however, may be helpful mainly to those more sensitive to the nuances of the language.

Test-Wiseness

All of this comes under the general heading of test-wiseness, a topic that has been talked about a lot but rarely studied systematically. *Test-wiseness* is the ability to use test and situational characteristics to improve test scores (Millman, Bishop, & Ebel, 1965). This does not imply trickery; the psychometric problem is mainly that errors are made when people lack test-wiseness and have substantially reduced scores because they missed cues others might perceive. Despite the lack of research, I think a test-wise person will, among other things:

1. Pay careful attention to instructions, oral and written, and be sure of understanding the task posed by the test and the basis for responses. A test-wise person will *not* expect simply to pick up insights into the nature of the test while taking it.
2. Ask for clarification of instructions if needed.
3. Begin working without delay and work steadily and as rapidly as possible without risk of misreading or clerical error in responding.
4. Read items carefully enough to be sure what is required.
5. Work rapidly enough to have time to check answers for errors in reading items or recording responses—and do so.

The list could be very long, but it would include little that has been empirically studied. I repeat the earlier call for more research on individual differences in test taking methods and their relative frequencies in different demographic categories.

NORM-REFERENCED AND DOMAIN-REFERENCED TESTING

Test scores are often *norm-referenced*, that is, interpreted relative to the scores of people in a comparison (norm) group. Whether a score is considered good or poor depends on the distribution of scores in the norm group. Figure 11.3 shows percentile ranks associated with raw scores in three hypothetical distributions. An examinee with a score of 12 has answered half of the items correctly. It is a magnificent score compared to those in Group C, better than more than 99% of the scores in that group. Compared to those in Group B, it is about average, neither very good nor very bad. It is not good at all—in the bottom quarter—in Group A, the group with the best set of scores.

Norm tables are rarely consulted in employment testing. Expectancy tables are more useful, but they too are norm-referenced, comparing

Raw Score	Percentile Rank in Group A	Group B	Group C
24			
23	99.9		
22	99.4	99.9	
21	97.7	99.6	
20	94.3	98.5	
19	88.4	96.4	
18	79.9	93.0	
17	70.2	88.6	
16	60.0	83.1	
15	50.0	76.9	
14	40.8	69.9	
13	32.3	62.1	99.9
12	24.6	54.0	99.2
11	18.1	45.5	97.2
10	12.7	36.7	94.1
9	8.5	28.2	89.3
8	5.4	23.9	82.6
7	3.1	16.7	73.9
6	1.7	10.6	63.2
5	.8	5.7	51.0
4	.3	2.5	37.5
3		.7	23.7
2		.2	11.9
1			4.0
0			.4

FIG. 11.3. Differences in interpretations of a given test score with different norm groups; a raw score of 12 is in the bottom quarter of the distribution in Group A, slightly above average in Group B, and outstanding in Group C.

candidates with each other. In a set of candidates, those with higher scores at any level are preferred over those with lower scores. Hiring the best of a poorly qualified lot, however, is poor management. In a test of prerequisite job knowledge, if every examinee should have a very high score, it is not helpful to say that someone with a very low score is less ignorant than a lot of other people and should therefore be chosen.

An alternative to normative interpretations was originally called criterion-referenced interpretation. In it, scores are interpreted relative to the content domain being tested; like Green and Wigdor (1991), I think it is more appropriately known as *domain-referenced interpretation*.[3] Under

[3]Not everyone shares this preference. Linn (1994) considered "domain-referenced" to require domain specifications too rigid to be feasible for any but extremely narrow, finite domains; he said that "criterion-referenced" refers to "broader, fuzzier, but more interesting achievements" (p. 13). Glaser (1994), who introduced criterion-referenced testing (Glaser, 1963; Glaser & Klaus, 1962), prefers the original term, pushing aside the barnacles of misinterpretations of his idea that occurred over the years.

either term, the basic idea is that a domain of accomplishments is identified and defined. It should be defined clearly enough that people, even those who disagree about the domain, can generally agree on whether a specified fact or achievement is in or outside of it. Measures of the domain should fit the definition, and scores should be explicitly interpretable in terms of it (cf. Hambleton, 1994).

In domain-referenced testing, the domain, not a point in a score distribution, is the criterion for referencing or interpreting an obtained score. A score of 12 on a 24-item test may mean knowledge of half of the content, but a better, fuller interpretation can identify the half not known. Content domains are rarely homogeneous, a fact permitting diagnostic uses of domain referencing.

Developing Domain-Referenced Tests

In theory, any test can be used for either norm- or domain-referenced interpretation. In practice, tests may be developed differently for these differing purposes.

Clarity of test purpose, always important, is especially so in domain-referenced testing (Popham, 1974, 1978, 1994). It is not enough to say that a test's purpose is to measure knowledge of repair procedures for electronic typewriters. Defining "knowledge of repair procedures" requires clarity about component content areas; components should be assigned relative weights, and the kinds of items to be used for each component should be specified (Popham, 1994).

Internal consistency, and the reasonable assumption of transitivity it permits, is a basic requirement in norm-referenced testing; people being compared should be compared on a common basis. It is less important in domain-referenced testing; in fact, if it is very high, the content domain may be too restricted. Of course, without some minimal internal consistency, scores have no meaning. Components of a content domain that are uncorrelated, or negatively correlated, should be separately scored.

Evaluating the validity of a norm-referenced test is primarily correlational, either in the sense of criterion-related validation or of confirming and disconfirming construct interpretations. Evaluating the validity of a domain-referenced test calls for expert judgment of the match of items to the specified content domain.

Mastery or Nonmastery

I began the discussion of tests with a domain-referenced example, the old oral trade tests. Cut scores distinguishing journeymen from apprentices or people in related occupations were interpreted as minimal levels

of content mastery. Setting a standard for designating mastery is judgmental, even when aided by empirical data. My general aversion to unnecessary cut scores applies here as well. Degrees or areas of mastery may be important for many personnel purposes. If so, such global designations as master or nonmaster may preclude useful flexibility and the diagnostic potential of a valid sample of a well-defined domain.

TESTS REQUIRING CONSTRUCTED RESPONSES

Some Criticisms of Traditional Testing

Critics of traditional testing find much to dislike. Some cognitive theorists complain that multiple-choice tests ignore thought processes. Some social critics consider them biased. Some educators agree with both and, further, blame testing for many educational problems. A general theme is that traditional tests, especially in multiple-choice form, are superficial measures of the wrong things. For some critics, the cure requires tests using free or constructed responses.

Contemporary educational thought seeks improved education for work readiness—education that teaches students to think clearly and creatively, produce ideas, evaluate information, and be cognitively effective—and it considers constructed-response testing to be the best way to assess success. American government initiatives (like Goals 2000, 1994) follow the crusade. The agenda is laudable, its perspective myopic.

An older tradition in employment psychology fits this new movement in salient respects. In it, standardized, objective, multiple-choice tests have coexisted with short answer tests, work samples, written statements of career objectives, problem-solving exercises, and other kinds of constructed responses. Three points should be clear before going on. First, "construction versus choice"[4] is not a distinction between new and old ideas. Second, freely constructed response and multiple choice are methods in a continuum from tests greatly constricting responses an examinee can make to those posing hardly any constraint at all. Third, the choice of a method should be based on the purpose, not on some all-encompassing merit supposed for one option over all others.

Levels of Constraint in Item Response

Bennett, Ward, Rock, and LaHart (1990) developed a taxonomy of item types (also presented in Bennett, 1993a). I follow it somewhat, but the presentation here is also influenced by other sources (mainly Snow, 1993).

---

[4]"Construction Versus Choice in Cognitive Measurement" is the title of the edited volume from which much of this section is taken (Bennett & Ward, 1993).

Multiple-Choice Items. Multiple choice and related item types anchor the maximum constraint end of the scale. Multiple-choice items differ. Some require only easy recognition; some require intervening construction in that the examinee must construct at least an approximate response before being able to make an informed choice among the options (Snow, 1993).

Selection or Identification Problems. These are also multiple-choice items, but quite different from the customary sort. An example (Bennett et al., 1990) is a passage of about 100 words, with several words underlined. Below the passage is an alphabetized list of more than 50 words, plus the phrases "no change needed" and "no appropriate replacement listed." Each underlined word is a "possibly inappropriate word choice"; the examinee is to decide if it really is inappropriate in its context and, if so, to pick a word from the list that would be more appropriate. There are far more options than a traditional multiple-choice item, so guessing is unlikely to help much. Moreover, it seems to call for mental reconstruction before searching for a replacement word.

Reordering or Rearrangement. The task here is to arrange items in a correct sequence; examples include solving anagrams, ordering sentences to make a logical paragraph, or ordering pictures to make a story, among others. Components can be arranged according to size, merit, complexity, chronology, or other principles.

Substitution or Correction. These items require the examinee to identify and correct a problem in material presented. In verbal materials, the problem may be a word that does not fit. The substitution comes not from a prepared list but from the examinee's own knowledge. Similar items could be nonverbal (e.g., wiring diagrams or abstract mechanical gadgets).

Short Answer Items. These typically require writing a word or phrase. It might be a word or phrase omitted from a written passage, a definition, a solution to a problem, or an answer to a question.

Construction or Production. These items require the examinee to produce something: a paragraph, a list, a graph, an architectural drawing, a gadget, an essay, and so forth. Work sample tests are common examples.

Oral Descriptions of Production Processes. These might include a teach-back procedure in which the examinee explains concepts, procedures, structures, or systems—usually but not necessarily orally.

Demonstration, Presentation, or Performance. These assessments require observing actual (or recorded) task performance. Auditions are obvious examples; others might be repairing a malfunctioning engine, diagnosing an illness, or giving a lecture. The assessee usually knows the nature of the required performance ahead of time and has time to practice it. Procedures might be standardized (e.g., all musicians playing the same composition), or the required performances may differ for different assessees.

Achievement Samples Collected Over Time. A person may be asked merely to develop and bring in samples of his or her best work. Standardization is truly minimal, "best work" being the assessee's own judgment. These usually are products of creative thought or skill, such as paintings, essays, short stories, recordings of performances, ad layouts—whatever is to be assessed. They ordinarily do not include intangible products like negotiated settlements between conflicting parties or reorganized production systems, but they might include written descriptions of the history of such problems, procedures used in working toward the outcome, and a reasonably objective evaluative summary of it. Note that this and the two or three methods just mentioned do not fit the definition of *test* given at the chapter beginning; they are neither standardized nor objective enough to score very reliably; they do not measure specific constructs, and the assessments may not even be ordinal in nature.

Scoring Constructed Responses

Only rarely can constructed responses at the extreme of the continuum be dichotomously scored as right or wrong, acceptable or unacceptable, safe or dangerous, workable or unworkable. The responses range widely from very poor through ordinary to excellent, even elegant; the place of a response in that continuum is a matter of judgment, often expressed as a rating.

Interrater or inter-scorer reliability is a problem. It should not matter which rater scores a response. For example, it should not matter whether it is scored early or late in a scoring session, or whether the test was taken on one date or a later one. Such unwanted sources of variance might be studied by generalizability analysis.

Some free responses can be scored reliably against a key or standard identifying acceptable and unacceptable responses. Partial credit might be given. A good (correct or acceptable) response might be given a score of 2, a poor one scored 0, and one in between scored 1. Items to be scored might be task components, whether an essay, work sample, or portfolio.

Performance on a driver's license examination may be a sum of scores on parallel parking, changing lanes, stopping, turning, and other parts.

The choice of an overall score versus component scores depends on the purpose. An overall score is useful if an either–or decision is to be made (as in selection or licensing decisions), but analytic scores are more useful for training. I think the analytic process can improve final score reliability in any case.

Scoring Keys for Essays. Keys can be prepared for essays. Those developing the scoring key may have in mind a list of content topics that should (or could) be included in a response to the stimulus question. A short rating scale (perhaps three levels) might be used in evaluating the quality of the coverage of each topic; the score may be the sum of the quality points awarded to the listed topics discussed in the written essay, perhaps giving more weight to some topics than to others. Such a key enhances scorer reliability. If the key includes all appropriate content, no one will come close to a maximum score; with even liberal time for completion, examinees are not likely to think of everything the content experts developing the key could include. A domain-referenced score of only half the possible score might be considered excellent performance.

Where possible, essays should be scored by two or more different people, each with the same understanding of the key and its use. If only one scorer is available, responses should be rescored at a later time, in a different order, without knowing the examinee's identity or earlier score. The process may improve reliability, but it does not show how to use the double scores. They might be summed or averaged; where discrepancies exceed an accepted level, some procedures require scorers to meet and reconcile differences, or a third scorer might contribute to the average.

Computer Scoring. New developments permit less freely constructed responses, scored dichotomously, or even completion items, to be scored using a computer or a scannable answer sheet. A word or number can be encoded on an answer sheet in a grid like those used for recording the examinee's name or identification number (Braswell & Kupin, 1993; Bridgeman, 1992). Computers may be feasible for even less constrained responses. A system developed by Bejar (1991) showed promise for scoring constructed solutions to architectural design problems, Braun, Bennett, Frye, and Soloway (1989) developed an expert system approach for problems on computer programming, and Bennett (1993b) incorporated artificial intelligence and expert systems in yet another proposal. Such scoring systems, of course, are not yet operational for most users.

Scoring Models. Whether items call for short answers or essays, or are scored dichotomously or with partial credit, scores on each item may be added to form a total score. Such scoring is a compensatory model,

probably useful for scoring end products like essays, typewritten letters, or welded assemblies. However, where testing objectives focus on process, conjunctive or disjunctive models may be more appropriate. A compensatory model may not even be feasible if process components are sequential.

An appropriate score on a set of work sample tasks may be the poorest score for any component. In another setting, for another job, the important consideration may be doing something well, so the score may be the best of all component scores. Or the scoring model may involve some combination of models. Conceivably, a work sample could include a variety of tasks, some of which are enough alike to combine. Within groups of tasks, compensatory scoring might be used, but the total score across those groups might be based on a much different algorithm.

Comparisons of Constructed and Multiple-Choice Items

Several studies have compared relatively constrained free response and more traditional multiple-choice testing. A common assertion is that multiple-choice items measure superficial constructs and that constructed-response items measure deeper constructs. That is, the two formats are said to measure different things. Is this a tenable hypothesis? Much evidence rejects it. Bridgeman (1992) reported little difference in the abilities measured when free responses are numbers or just a few words. Ward (1982) showed that open-ended forms of some verbal aptitude item types can be as reliable as multiple-choice items and that they require only slightly greater time limits. Testing the hypothesis that multiple choice and free response define different verbal aptitude factors, he found only a single factor—no evidence of different constructs. After comparing a 50-item multiple-choice test in computer science with five essay components in the same test, Bennett, Rock, and Wang (1991) also reported a single factor, concluding that the evidence did not support the stereotype of multiple-choice and free-response formats as measuring different levels of ability. In general, the rhetoric of the constructed response movement is not empirically supported.[5] Even the least constrained response formats seem unlikely to measure constructs differing greatly from those measured by the most constrained response formats. It seems more likely that they have more contaminating sources of variance, diluting validity (Snow, 1993).

For any measurement method, it is important to formulate carefully, and investigate systematically, rival hypotheses about constructs measured. One contaminant commonly invoked is test anxiety; its effect is

[5]Bennett (1993a, p. 2) attributed to C. R. Reynolds the motto, "In God We Trust; All Others Must Have Data." It seems fitting here.

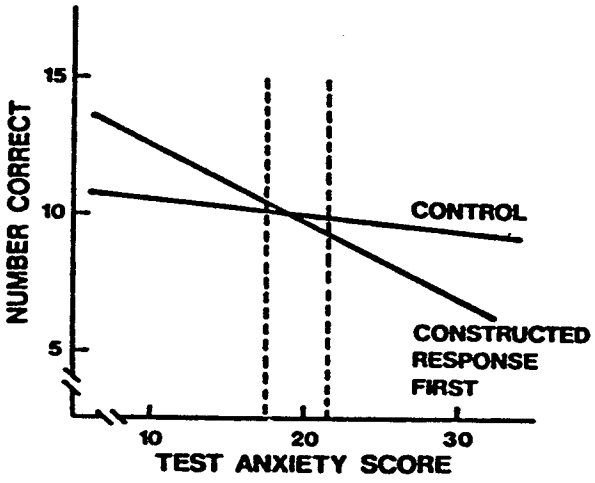

FIG. 11.4. Regression slopes of achievement test score on test anxiety score where (a) conventional multiple-choice format is used and where (b) responses were first constructed and then recorded as multiple-choice responses; differences between the dotted lines are nonsignificant. From Snow (1993).

shown in Fig. 11.4, adapted by Snow (1993) using data from Schmitt and Crocker (1981). For the typical multiple-choice format, test scores were slightly negatively related to test anxiety. Different instructions, telling the examinee to construct a response before choosing one of the options, produced the more severely negative slope. The constructed response form includes more unwanted variance due to anxiety.

Multiple-choice items typically yield more reliable scores than do even similarly constrained free-response items; Wainer and Thissen (1992) estimated that it cost $30.00 to get enough testing time and items in the free-response format to equal the reliability per penny for multiple-choice items. Martinez (1991) found multiple-choice items easier but somewhat less discriminating (i.e., lower biserial *r*s), but constructed-reponse items were more likely to be omitted. Student examinees (and, maybe, candidates for employment opportunities) prefer multiple-choice items but think open-ended questions measure abilities better (Braswell & Kupin, 1993; Shepard, 1991).

A Tentative Conclusion

There is much hope but little evidence that constructed-response testing is superior to more constrained responses. The rhetorical excesses may have inhibited research. Some rhetorical arguments may be worth pursuing—for example, a distinction between the traditional concern with the examinee's level of proficiency and what Mislevy (1993) called the *architecture of proficiency*:

> The ascendent view [i.e., the constructed response view] originates *from a perspective more attuned to instruction than to selection or prediction.* Learners increase their competence not simply by accumulating new facts and skills . . . but by reconfiguring knowledge structures, by automating procedures

> and chunking information to reduce memory loads, and by developing strategies and models that tell them *when and how* facts and skills are relevant. (Mislevy, 1993, p. 75, italics added)

And,

> Rather than seeking long-term, stable characteristics that are immune to change, a test in this context is meant to provide information about characteristics of an examinee that are *ripe for change. The problem of interest is one of diagnosis* or optimal assignment to instruction; the decision is viewed as shorter term; the options are cast not in terms of level of persistent proficiency but of *architecture of current proficiency.* (Mislevy, 1993, p. 79, italics added)

Forget the straw-person reference to characteristics immune to change. Ignore also the education-specific reference of the quotations. What is left is a potentially important distinction between level and structure of performance achievement. One hypothesis might suggest that a person who has moderate if not particularly high levels in several components of proficiency, and can be readily trained in others, may be a better bet for growing requirements for flexible personnel than one who has demonstrated a high level of proficiency within a much more limited set of procedures. I do not know of any test of the hypothesis, but it is testable. If it is supported, the diagnostic value of contructed response testing may have much practical value for personnel decisions.

Some constructed response testing may have further advantages. Consider again the model of personnel decisions suggested by Dunnette (1963). If two candidates present themselves for consideration for the same job, and if one of them claims several years experience on similar jobs and the other claims no more than ability and willingness to learn, the Dunnette model suggests that these candidates be assessed in different ways for different characteristics. The latter should be assessed on relevant stable, long-term aptitudes; a work sample or job knowledge test can assess the proficiency claimed by the former.[6] If concern for the architecture of proficiency does no more than encourage a reappraisal of the Dunnette model, it will have proven worthwhile for HR assessment needs.

There is surely something valuable for employment testing in the resurgence of interest in constructed responses. However, people doing measurement and assessment for personnel decisions must not be stampeded into their premature adoption, either by academic rhetoric or by

---

[6]The legality of this sort of differential testing has yet to be determined, but in principle it seems consistent with the ADA concept of accommodating those who are differently abled.

government activities that urge uncritical transfer of educational assessment ideas to employee or job candidate assessment.

PERFORMANCE TESTS

Performance testing in the workplace means assessing proficiency in some aspect of job performance. Performance tests may be cognitive or noncognitive, paper-and-pencil or "hands-on," and anywhere from the most to the least constrained kinds of responses. They may be criteria or predictors intended to predict no further than the immediate future. An applicant who does well on a welding test may be expected do good welding the first day at work; situational variables like equipment, materials, supervision, coworkers, or personal traits like motivational level, may determine whether a good beginning is continued. Although prediction is always implied, performance tests are used mainly to assess proficiency, skill, or knowledge at the time of testing—here and now, not at some future time. Unlike low aptitude candidates, those lacking knowledge or skill may acquire it through special training and reapply when ready.

Performance tests can be used (a) to predict performance on a higher level job requiring similar kinds of proficiency, (b) to identify outside candidates who need no training beyond a general orientation, (c) to identify training needs, (d) as a criterion in validation, (e) to provide proficiency-related interpretations of predictors, and (f) in performance evaluation. Only the first of these has a strong future orientation; the principal orientation of all the other purposes is here and now. Use as a criterion should be more common than it is, but its value as a criterion can be overstated. Performance testing usually describes how well tasks can be performed when the person is doing his or her best. Where testing is intended to predict actual performance, not a hypothetical maximum level, performance test scores may be inappropriate criteria. Again, the method of assessment should fit its purpose.

Work Samples and Simulations

The most common "hands-on" performance tests may be work samples. They are well-established as predictors. Their criterion-related validity is consistently shown in reviews (e.g., Asher & Sciarrino, 1974; Cascio & Phillips, 1979; Robertson & Kandola, 1982; Schmidt, Greenthal, Hunter, Berner, & Seaton, 1977).

A work sample test is a standard sample of a job content domain taken under standard conditions. Aspects of the work process, the outcome, or both may be observed and scored. In a flight test for a pilot's license, the focus is on process; a check pilot has a checklist of required maneuvers and

evaluates how well each is performed. A candidate for an office job may be given a typed manuscript with many scribbled changes on it, be seated at a word processor, and told to prepare final hard copy; perhaps only the result is observed and scored. In either case, the work sample is a *standardized abstraction* of work actually done on a job. There are degrees of abstraction. A work sample might be faithful reproductions of actual assignments, sanitized simulations of critical components, or the extreme abstraction, here and now measures of isolated skills used on the job.

Simulations imitate actual work but omit its trivial, time consuming, dangerous, or expensive aspects. They may imitate a task almost exactly, as in some simulations of aircraft cockpits. They may imitate only the general flavor of reality, as in assessment center management exercises.

Other possibilities carry abstraction still further. Performance tests might use *talk-through* interviews (Hedge, Teachout, & Laue, 1990) to describe the steps, tools used, and decisions made in doing the job. A work diary might be used. A collection of product examples (a "portfolio") may be evaluated. Even a multiple-choice test may abstract from overall performance the knowledge and understanding of processes, tools, and choices that make up performance on the job. Simulations that are not highly abstracted are known as *high fidelity* simulations; the greater abstractions may be *low fidelity* simulations (Motowidlo, Dunnette, & Carter, 1990).

All of these are performance tests. In all of them, components, tasks, or required behaviors are drawn (i.e., abstracted) from overall job demands. They may be manipulative, sensory, or cognitive; they may be faithful, obvious samples of work done or abstractions recognizable as samples only by knowing the logic leading to their use. They may test knowledge, understanding, or skill. In short, they vary widely in nature, in content, and in fidelity of imitating real job performance.

Most people seem to assume that the more faithful the simulation, the better it will predict future performance. As Sportin' Life in Porgy and Bess said, "it ain't necessarily so." Prediction may be better when faithfulness of simulation is set aside to focus on the more enduring essence of a job. Low fidelity simulations have been good predictors when they represent the essence of both the job and the simulation (Motowidlo et al., 1990; Motowidlo & Tippins, 1993). Perhaps the ultimate abstraction was reported by Arnold, Rauschenberger, Soubel, and Guion (1982) who abstracted simple arm strength from steel mill labor jobs. The arm dynamometer test was not a faithful imitation of any real task, but most of them demanded arm strength, so it measured a critical aspect of job performance.

Developing Work Samples. Work sample development begins with job analysis, although not everything the analysis identifies is included. Distinguishing a "universe" from a "domain" (Lawshe, 1975), I described choices in developing tests of content samples in four stages (Guion,

1979). A complete job analysis identifies a job content universe. The part of the universe to be assessed is a job content domain. Related assessment possibilities (including scoring methods) make up a test content universe, and the choices among them define the intended test content domain.

Proficiency is the construct measured by a work sample, but it takes many forms. For a criterion, it should identify all tasks critical for overall performance. For selection, it omits critical tasks learned on the job. Ordinarily, tasks defining proficiency should be those that many, but not all, examinees are likely to perform well. Most work samples use only frequent tasks; rarely performed tasks might be in the domain to identify those who can handle unusual job situations.

Equipment or material used should match that actually used on the job (instead, as so often happens, of stuff not yet thrown away). Tolerances and procedures for monitoring equipment should be established; if holes into which things are inserted get larger over repeated testing, monitoring hole size may be an important aspect of standardization. As always, pilot studies should evaluate the clarity of instructions, scoring procedures, and characteristics of test components (e.g., items) as well as overall reliability and validity of scores.

Work Sample Scoring. Scores are usually ratings. An overall rating of process, product, or component part can be dichotomous (e.g., satisfactory or unsatisfactory) or a scale point. A work sample product might be matched to one of a set of samples previously scaled from very poor to excellent (Millman & Greene, 1989); the score being the scale value of the sample it most closely matches. More objective measures can be used. A score on machine set-up might be the time required to do it. The score can be the pounds of pressure required to break a weld. A computer might count the number of corrections made in a sample word processing task. Ratings predominate, however, and their associated problems (see chapter 12) can be helped with procedures like these:

1. Job experts should choose work sample content, specify desired performance, and provide at least a preliminary scoring key or protocol.
2. Scorers should be trained to use the protocol: what to look for and how to evaluate specific events or product components.
3. The same performance or product should (if possible) be evaluated by two or more independent observers; impermissible differences in ratings should be defined and the procedures for reconciling differences prescribed.
4. All possible procedural safeguards of reliability should be built into the scoring system.

Evaluation. Evaluation of a completed work sample test follows the set of questions given in chapter 5. Guion (1996) restated these questions explicitly for performance tests; three further questions, drawn from Mehrens (1992), should be added for performance assessment:

1. Are the content domains sufficient for the purposes declared? "In general, performance assessment measures a narrower domain than multiple-choice testing but assesses it in more depth. Is this good?" (Mehrens, 1992, p. 7).
2. Are domains defined tightly enough? Do experts agree on whether a given test component belongs in the domain, as defined?
3. Is there evidence that performance on the work sample generalizes to the larger content domain?

Noncognitive Performance

Physical Abilities. Measuring strength, muscular flexibility, stamina, and related abilities usually requires equipment and individual testing. Equipment needs described by Fleishman and Reilly (1992) are often simple. Assessing stamina may require an electronically monitored treadmill with an accompanying electrocardiograph, but a simple step test can also assess stamina, although with less precision.

Fitness Testing. Task performance, physical fitness, and health may be related. Task performance, as measured by a work sample, may be supported by physical abilities (e.g., stamina). Abilities are supported by biological systems (e.g., cardiorespiratory systems), which may be impaired by health problems. A person with emphysema suffers cardiorespiratory impairment, resulting loss of stamina, and difficulty in tasks like climbing stairs. Poor fitness is a problem for both the person and the employer.

Medical and physical testing should have higher than typical priority, if for no other reason than protection from litigation. Litigation can spring from many directions (including getting hurt in fitness testing). An organization may be legally liable for hiring unhealthy or physically inept employees (under the concept of negligent hiring); there is an opposing liability for discrimination against the disabled. Employees who hurt themselves or develop health problems because of physically demanding jobs add to worker's compensation costs. Performance errors or accidents stemming from fatigue or clumsiness may bring suit from fellow employees, customers, or the general public. Rejected or underplaced applicants may sue under civil rights laws.

The potential cost is too great, both in the risk of litigation and the risk of physical pain, to continue using arbitary, poorly assessed standards of fitness or physical skills. Some perennial questions must be faced. For example, a physically demanding task may not be performed often but, when it is, injury might result. Should employment decisions be based on the ability to perform that task? Because of its infrequency, a worker may have little opportunity on the job to develop or maintain the necessary physical skill. On the other hand, infrequency may give time for rest and recovery between occasions. Should the job be redesigned, with the rare but risky task assigned to another job with similarly demanding tasks regularly done? Or should such tasks be spread around? Sometimes there is no option. In police work, sometimes defined as boredom occasionally interrupted by panic, the need to meet unusual physical demands is always present. Should physical fitness testing look at the job as a whole or at its maximum requirements? Should it be assessed periodically?

How much loss of musculoskeletal flexibility, cardiovascular impairment, or hearing loss must be experienced before job performance deteriorates? In medical examinations, answers to such questions are usually left to the judgment of the individual physician—but they are rarely validated. Fleishman (1988) offered a promising approach to systematizing such judgments. A guide to impairment evaluation published by the American Medical Association (1977) was tied to his scales for analyzing physical job demands; guides were developed for physicians to use in determining whether a given level of impairment would prevent effective performance of specific job tasks.

In many jobs, recurring personnel decisions may be made almost daily on employees' here-and-now readiness to work; for example, is this pilot fit to fly today? Is there an impairment that would make this construction worker's job especially dangerous today? Temporary proficiency impairments may be due to medication or drugs (including alcohol), fatigue, illness, or preoccupation with nonwork stresses. Drug testing is increasingly widely used, but drug tests or tests for blood alcohol level or body temperature do not assess impairment. It may be more useful to use performance tests of the specific proficiencies required, or perhaps physiological measures of performance impairment. Rizzuto (1985) found that even a mild dose of Valium resulted in slower neurological transmission and deficits in visual acuity, attention, and intensity perception; he found performance deficits in a visual tracking task using neurological measures of evoked responses that also identified the processing areas affected.

Olian (1984) offered a unique suggestion to reduce health hazards genetic testing for people expected to work in environments where they risk disorders stemming from specific hazards (e.g., hazardous chemicals) some people, genetically, may be at higher risk than others. Her sugges-

tion deserves consideration; as Murphy's Law says, where something can go wrong, some day it will, and the possibility of harm to people especially sensitive to a given hazard is real. Maybe one reason why the suggestion has not been considered further is the justifiable paranoia employing organizations experience when trying something that has not yet been tested in court.

Sensory and Psychomotor Proficiencies. Work combines cognitive, muscular, sensory, and attitudinal components; a useful work sample might focus on the sensory component. Requisite here-and-now job performance may include sensory proficiency such as correct identification or distinctions of distant shapes, colors, musical pitch, or unseen but touched objects. Except for some classic studies (e.g., occupational vision; see Guion, 1965; McCormick & Ilgen, 1980), little research has addressed the assessment of sensory skill for personnel decisions. Fleishman and Reilly (1992) identified assessment methods for a few sensory abilities; more importantly, perhaps, they identified some important skills (e.g., night vision) for which no existing measures were identified; these, too, are ripe areas for research.

Psychomotor skills, especially dexterity and coordination, are more widely tested. Especially common is the use of dexterity tests, usually requiring examinees to insert pegs or pins in holes, as in Fig. 11.5. Scores can be the number of pins (or assemblies) inserted within a time period or the amount of time required to fill the board.

Examples of tests for other psychomotor skills are provided by Fleishman and Reilly (1992). Commercial psychomotor tests are available, but sometimes manipulations imitating those required on a job should form

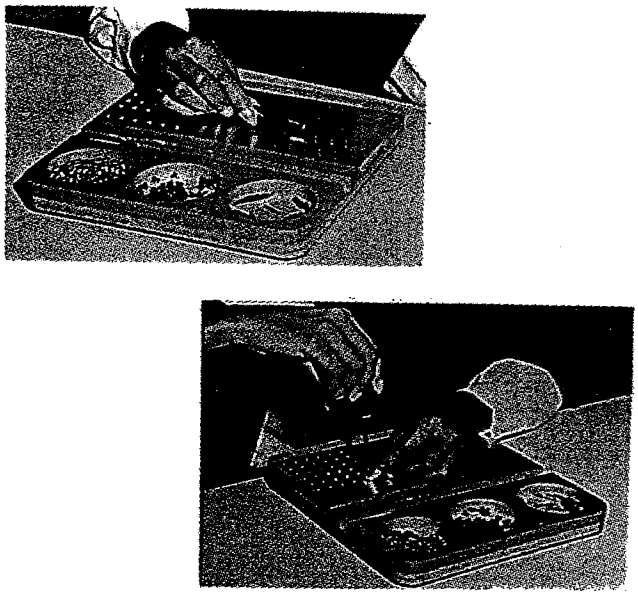

FIG. 11.5. Testing dexterity with the Crawford Small Parts Dexterity Test. From Guion (1965).

the test. Job analysis can identify the recurring stimulus patterns and the kinds of coordinated responses required.

High skill levels in some sensory or psychomotor areas may compensate for deficiencies in others, in work as in more general life skills. The compensatory development of unusual auditory skills among the legally blind is one example; the extraordinary skin and muscle sensitivity of the deaf and blind Helen Keller is legendary. Examples need not be so dramatic to have implications for personnel management. Rehabilitation counselors tell about people lacking certain sensory (or motor) skills performing well on jobs many employers would have denied them. Hope for finding compensatory skills is based more on anecdotes than on research. Evidence does not yet lead to general propositions about genuinely compensatory patterns.

Assessment of Basic Competencies

It is hard, and maybe not useful, to distinguish between ability and competency, but more and more frequently, calls are made for testing competencies rather than general abilities (but see Barrett & Depinet, 1991). Some distinctions may be sensible. *Competency* may refer to here-and-now performance, *ability* to aptitude for future performance. Competency scores are likely to be domain-referenced, ability scores to be norm-referenced.

There is much contemporary talk about basic competencies, especially those required in "workplace readiness" (cf. Resnick & Wirt, 1996, p. 21). I propose defining *basic competencies* in employment as the acceptable performance of simple things a person must do on a job, things like adding whole numbers, reading simple instructions, writing notes on problems or activities, or reading blueprints, among other examples—things an employee on a particular job may be expected to do without help, instruction, or accumulated job experience. The nature and complexity of a basic competency, so defined, differs for different jobs. A cashier must count money accurately. So must a pizza deliverer, who must also drive a car. An electrician must read wiring diagrams. An office clerk must read, alphabetize, and maybe type. These are basic competencies, defined by these jobs, not for work in general.

"Authentic" Performance Assessments

The phrase, "authentic performance assessment," currently popular in education, known also as *performance-based assessment*, has spilled over into legislation and regulation. According to some (e.g., Barton, 1996), its use was mandated in Goals 2000: Educate America Act (Goals 2000, 1994).

Saying it is a mandate is debatable, but Section 207(a) of the Act calls for "alternatives to currently used early childhood assessments"—a phrase reflected in other parts of the Act as well—if they are "valid, reliable and consistent with relevant, nationally recognized, professional and technical standards" (Goals 2000, 1994, Sec. 211[5][B]). The concept of authenticity has been defined in many ways, not always consistently. The common thread may refer to assessment based on physical and cognitive challenges actually faced in the school (and, by extension, the job; cf Linn, 1994). Performance assessment always implies that someone has done something evaluated by someone else; in that sense it is hard to imagine an inauthentic assessment, even if an invalid one is quite easily imagined. Authenticity seems to refer mainly to the match between the stimulus content of the assessment material and the stimulus content of performance in the classroom or job site. If this is correct, authenticity in performance assessment is little more than a buzzword for fidelity of simulation. In that sense, the degree of authenticity may be inversely proportional to the degree of abstraction. As in other testing, authentic performance assessment should not be used when the assessment procedure can result in harm to the assessee, others, equipment, or the services the assessing organization provides. Lower fidelity assessment may be more valid and less dangerous. Work sample, simulation, and special skill testing are all performance measures, successsively more abstract perhaps, but not certainly, the less abstract ones may more authentically match critical job stimulus content.

Probationary Assessment. Authentic assessment of job performance should assess day in–day out performance; maybe the test is performance in probation. It may be authentic, but it has pitfalls. Probationary tasks that faithfully sample later assignments surely provide a representative, job-related, authentic job sample. However, in assessment, inferences are drawn from evaluative scores (usually ratings), not assignments, and validity depends on those inferences. Valid inferences about performance can be based on less than fully representative probationary work; a highly representative set of assignments can be spoiled by invalid evaluations.

Records and Portfolios. A current icon of authenticity is portfolio assessment. A *portfolio* is a performance record consisting of documents attesting to performance, descriptions of performance, or products of performance. It can include indirect reflections of real-world performance such as awards, production records, or commendations or disciplinary actions. A letter to an interstate trucking firm commending a driver for an exemplary action is an observation and evaluation of the driver's work; a record with several such letters may be evaluated more highly than a

file with several letters of complaint. In a quite different sense, job-related biographical data such as achievement records may find a place within a portfolio. A supervisor's diary of observations of employee performance—an *incident file*—can be in the record (Guion, 1965). All of these have been used in performance evaluation, but they are more likely to be called an employee's personnel file than a portfolio.

The portfolio concept seems different. A portfolio may be assembled by the person whose performance is being evaluated. It is likely to consist of content that is best rather than representative, so it assesses maximum, not typical, performance. An assessee may choose examples of his or her best work and submit them for critique and evaluation. Note the sequence of evaluations. The first is the assessee's, who uses a personal conception of excellence in assembling the portfolio. Later a teacher or HR specialist evaluates the quality of the items collected. The disparity in concepts of excellence introduces an uncontrolled source of variance.

Portfolio assessment may help in some employment decisions. If several managers are being considered for a promotion to the executive level, each of them may be asked to submit reports on two or three programs he or she has initiated and considers noteworthy managerial achievements. In choosing people for so-called "talent" jobs, (e.g., writing advertising copy, designing consumer products, etc.) portfolios representing relevant prior work might be required. The manager of a dinner theatre might ask for a portfolio of prior stage experience before deciding which applicants for a new play will be auditioned. An investment counselor might be asked for a portfolio of portfolios. These are not commonplace jobs, but principles of assessment for personnel decisions ought not be limited to the ordinary. It may be instructive in such jobs to evaluate assessees' ideas of the best.

Enthusiasm for portfolio assessment is largely due to dissatisfaction with traditional tests. It is inappropriate, however, to criticize traditional testing by one set of rules while justifying a preferred approach by different rules. Authentic performance assessment needs to satisfy, minimally, some common evaluative requirements. Foremost among them is that the assessment should accomplish its objectives. An assessment program may be intended to shed light on how well the person assessed can do what is expected in a curriculum or a job. It may be deficient if too much of the total domain goes unassessed (Mehrens, 1992).

Second, there must be *some* standardization. When people are compared, the comparisons should be (in the overworked cliché) on a "level playing field." When assessment of different people is based on different kinds or levels of performance demands, the assessments may not be comparable at all. Standardization sometimes seems to be a dirty word to some people, but it is essential for fair personnel decisions in which some candidates

come out ahead of others. Standardization may be a dirty word, but procedural justice is not. In the litigious climate in which employment decisions are made, the playing field of unstandardized assessments seems legally dangerous, quite apart from its contribution to unreliability.

Reliable assessment is the third requirement. Koretz (1993) reported low interrater reliability coefficients for the Vermont portfolio project, ranging from .33 to .43, although later estimates were somewhat improved (Koretz, Stecher, Klein, & McCaffrey, 1994). With experience, of course, scoring procedures will be improved and so will reliabilities. Reckase (1995) has shown analytically that acceptable reliability is possible. But it will require a great deal of work.

Finally, cost should be considered. "Authentic" assessment is extremely expensive. Providing equipment for simulation, for example, can cost more than the working equipment it simulates. The costs of finding, training, housing, and paying essay or portfolio readers are far greater than the costs of more traditional programs. The evaluative question is both psychometric and economic: Is the assessment benefit worth the cost? There is no general answer for all occasions; benefits and costs must be estimated in the light of specified purposes, conditions, and alternatives.

ELECTRONIC TESTING

Technological change can make tests obsolete (e.g., stenographic tests used circa 1940) or create opportunities. Electronic technologies offer new ways to do conventional testing and new ways to do unconventional testing.

Motion Picture and Videotape Tests

In large public jurisdictions, several thousand candidates might be tested simultaneously in different locations with different test administrators. Beyond logistics challenges, mass testing may pose psychometric problems only for tests requiring rigid time limits; differences among examiners in timing accuracy is a source of error. Even appropriate speeded tests are avoided because a timing error may unfairly disqualify or give unfair advantage to examinees in at least one location.

Putting a test on film, videotape, or slides, projecting items under controlled conditions, can solve the practical problem of controlling instructions and time limits. There are other psychometric advantages as well. By controlling the time for individual items, rather than an overall time limit, all subjects can attempt all items, and internal consistency analysis is feasible. Individual item characteristics can be changed, such as changing item difficulty by changing the item's exposure time.

A movie was the medium for a test of perceptual skill for the selection of police officers in a large midwestern city.[7] Perceptual accuracy is important in police work, but the police skill differs from that usually tested. An officer may look for something specific, such as a license number; more often, the officer must simply be alert to things that merit curiosity. A police officer on patrol or doing investigative work must attend to detail without knowing which details might be important. Perception must be accurate, and it must be quick.

The first section of the film gave instructions and some illustrative but very brief scenes narrated by an off-screen voice; these scenes were the basis for a subsequent memory test. The rest of the film showed several brief episodes, each followed by a set of multiple-choice questions about it. Some scenes had a story line; some were simply camera shots, for example, of people enjoying themselves on a summer day at the park. The format permitted use of behavioral stimuli (Guilford, 1959) that could not have been included in traditional tests, but it was otherwise fairly traditional. As in any other test development, we should have built the test in stages, basing each stage on pilot study data from the one before. For each scene we should have had pilot studies for item analysis and have chosen the best items; we did, perfunctorily, but we did not do an analogous pilot to choose scenes. The script was written, and the film was shot, edited, assembled, and declared a take-it-or-leave-it test. Time pressure and economic constraints forced a development process that was less than optimal. Nevertheless, construct-oriented studies led us to conclude that the scores measured a "perception-on-the-run" construct; they disconfirmed both plausible contaminants and the expected independence of three component constructs. Several criteria were predicted as well from the movie scores as from tests with more nearly complete research backgrounds.

Some Video Tests. I described the movie test because of firsthand knowledge of its construction and evaluation and because its research flaws could be described without embarrassing others. Videotaping is more up to date and has the further advantage of permitting more experimentation with alternative scenes and scripts.

Video tests have been reported for assessing situational judgment in customer service jobs, among others, with gratifying validity coefficients (Curtis, Gracin, & Scott, 1994; Dalessio, 1994). These tests generally use the vignette and question approach; sometimes the effectiveness of depicted behavior is rated. A video test to measure work habits and team

[7]Kenneth M. Alvares and I did the research reported here; it has not been previously published. The work was quite extensive, but it was done under court jurisdiction and could not be published until the case was settled. I believe the case, like an old soldier, just faded away.

skills showed interpersonal problem situations typical of those that can arise in a factory setting—differences in adherence to work standards, responsibility, and interpersonal conflict. In pilot studies, correlations with traditional ability tests were nearly zero, but correlations with a combined productivity–quality scale and with contextual criteria (ratings of communicating and solving problems and work habits) were significant. Corrected for criterion unreliability, the validity coefficient of the video test was virtually identical to that of a composite of five predictors (the video test and four standard ability tests), optimally weighted![8] Except for significantly higher mean scores for Asians, no statistically significant mean differences were found among ethnic groups or between two age groups; women did somewhat better on the average than men.

Computerized Testing

Medium Effects. Early computerized testing (computer-based testing, or CBT) did little more than put the items of a traditional test in a computer, presenting the items one at a time, getting a response (perhaps with a built-in time limit), and moving on to the next item. CBT tests usually differed from paper versions in that (a) items could not be skipped and tried again later, (b) time limits were set for items rather than for the tests as a whole, and (c) scores could be reported immediately. Another difference was physical; printing methods gave the paper versions a readability advantage over early, relatively low-resolution monitors. It is worth asking whether such differences matter in test characteristics such as mean scores or variances or the constructs measured.

Because computer use was especially important in clerical work, Silver and Bennett (1987) thought the pre-eminent Minnesota Clerical Test should itself be computerized. Two versions using computers were developed. One was a simple shift from paper to computer; the other, used as a criterion, put the original left-hand column on paper and the right-hand column on the computer—requiring the back and forth focus between hard copy and screen that characterizes many data entry and data checking jobs. The hypothesis that CBT tests were necessary for computerized jobs was not supported; validity coefficients were not significantly different for the paper test and the computer translation of it. Significant mean score differences, however, suggested that the computerized version was more difficult. Moreover, correlations were not high enough to suggest identical constructs. Other studies have reported similar findings (e.g., Green, 1988; Mazzeo, Druesne, Raffeld, Checketts, & Muhlstein, 1991).

---

[8]Further information may be obtained from Dr. David P. Jones, President, HRStrategies, Inc.

A meta-analysis reported by Mead and Drasgow (1993) found the computer versus paper effect moderated by test speededness. Where a test is essentially a power test (i.e., time limits are generous enough that nearly all examinees can finish), the mean correlation (corrected for attenuation) between scores obtained with the two media was .97. However, highly speeded tests did not correlate as well across media; the mean correlation for speeded tests was .72.

Bell Atlantic's Universal Test Battery. If a paper-and-pencil test is working well, there is little point in fixing it with a computer. If the special advantages of computerized testing make it the medium of choice, one may as well develop a new test, and evaluate it, using the computer format from the start. A wide range computerized test battery developed for Bell Atlantic has been used for more than 100 nonmanagement jobs (Hough & Tippins, 1994). Traditional test batteries, well-validated, were available for many of these jobs, but the program was administratively burdensome.

A single test battery was developed for administration on laptop computers. It consisted of nine timed cognitive ability tests and a six-scale, virtually untimed, noncognitive inventory (with an added response validity scale). It required about 2 hours to complete. All candidates for hire or transfer were to complete the entire battery although only parts of it were expected (and found) to be valid for any one job. This feature permitted placement and job counseling as well as selection and transfer decisions. The tests were essentially conventional except that they were developed, analyzed, and validated in computer form. The system used modem connections with the central office mainframe, permitting scoring to be centrally controlled regardless of the remoteness of the testing location. Computerization also permitted computer-generated reports for examinees.

Advantages of Computer-Based Testing. CBT is not necessarily superior to traditional testing; the Silver and Bennett (1987) experience shows that translating a perfectly good paper test into a computerized one may be a waste of resources. A beginning testing program, however, might well be governed by the advantages of computer forms. Among them:

1. Testing conditions are more completely standardized. Variations in clarity and manner of speaking are not controlled when different examiners present instructions or test items. Most of all, readiness of examinees to begin is not controlled by the usual "Are there any questions?" routine. CBT instructions can, as on paper tests, provide samples of what is to come; on the computer, however, incorrect responses to sample items can delay the start of the test until a sequence

of sample items, developed to address differing problems in un
standing, is handled satisfactorily. That such care in developing inst
tions is rarely used does not remove the possibility from the lis
potential advantages.

2. Examinees usually enjoy CBT more than more traditional for
One was quoted after taking a computer adaptive test as saying, '
faster, it's funner, and it's more easier" (quoted by Green, 1991, p. 2

3. Johnson and Mihal (1973) found lower CBT mean differences
scores of Black and White examinees. They suggested that novelty a
the reduction of negative expectations might account for the findin
hypothesize that computer presentation reduces variance due to test-t
ing strategies.

4. Different kinds of items can be used. Computers can present changi
visual and auditory stimuli. As just one example, a mechanical knowlec
test could be devised with items showing equipment in motion w
questions about forces, problems, or errors in the graphic display.

5. Different, but potentially important, constructs can be measure
For example, response latency can be recorded, either to measu
something akin to reaction time or as an item characteristic to
considered in scoring. Where special skills are used, mean respon
latency toward the end of the test can be compared to mean latency
the beginning to measure learning during the test.

6. Programs can be developed that permit consultation of referen
materials (e.g., dictionaries, procedures manuals), in turn permitting iten
posing more complex problems. A computer-based simulation of arch
tectural practice used two monitors (Braun, 1994). One provided acce
to resources: excerpts from standard reference materials, prints an
drawings relevant to the test projects, and a "file cabinet" with project-rel
vant written material one might find in an office filing cabinet. The othe
monitor represented the architect's work place where the examinee doe
design work according to the task posed in a project vignette. The examine
can access either monitor at any time with simple mouse use.

7. Test taking strategies can be studied on computerized tests, an
perhaps scored and considered in interpreting the trait scores.

8. Test security is easier to maintain. A few computer disks can b
held secure more easily than can a few hundred printed tests. Moreover
the order of items in the sequence can be scrambled.

9. Item banks can be created, calibrated according to stable item
characteristics (either those of classical test theory or IRT), from which
computers can draw items according to specifications to make up unique
test forms for each examinee, permitting a large number of psychomet-
rically equivalent forms to be generated from the bank. Item banking

therefore offers a potential advantage for both test security and the common problem of retesting. Two different candidates may see *some* common items, but item differences would be substantial enough to reduce the test security problems associated, for example, with item memorization (Bergstrom & Gershon, 1995; Gibson & Weiner, 1996; Vale, 1996).

There are, of course, disadvantages as well. One is cost—more in programming than in hardware. Another is examinee computer anxiety, either because of unfamiliarity with computers or with the type of software used. However, these disadvantages seem to be decreasing over time.

There are special problems of standardization. A full keyboard can be daunting to people unaccustomed to computer use; special keyboards are often provided to match response needs. All examinees should be provided the same basic keyboard configuration—movable to make it maximally convenient for both left- and right-handed users—or alternative response methods (e.g., touch-sensitive screens, mouse pointers, etc.). Programs should be the same, with the same hardware demands, for all candidates. Displays should have the same resolution, color, and other features. All testing stations should use the same make and model of microcomputer. These and other considerations are described in detail by Green (1990). The speed of computer obsolescence exacerbates ordinary standardization problems.

Computerized Adaptive Tests (CAT)

Conventional testing is also known as *linear testing*; all items are presented one after another to all examinees. A high ability person flies through the easy items; only hard items show just how able that person is. Linear testing is therefore an inefficient use of testing time.

Adaptive testing, on the other hand, uses a branching algorithm and, therefore, fewer items. It begins with one item of moderate difficulty; the next one chosen depends on the response given to the first one—and so on until a predetermined criterion for stopping the test has been reached. If the first item is answered correctly, the next one may be more difficult. If the next one is answered incorrectly, the third item may be between the first two in difficulty. Adaptive testing has long been used in individually administered ability tests, but it required the combination of modern computers and the development of item response theory to bring it to its current level of sophistication.

Use of IRT in CAT. When a large set of items is stored in the computer, each with the parameters of its item characteristic curve and information function, a first item can be one of several with a moderate, mid-range

difficulty level. If it is answered correctly, the next item is harder. Wher it is answered, correctly or not, the information function of the two-item set and the person's ability level can be estimated. A third item can be chosen, based on its information function, that measures ability at that level with the greatest precision. The combined information from the three items provides a new, more precise estimate of ability and the basis for choosing the next item. This continues until the person's ability estimate does not change, or changes only within a narrow range according to a prescribed stopping rule, as shown in Fig. 11.6. The score is not the number of questions answered correctly; it is the estimated ability level. CAT abandons the idea of a standard set of items but not the idea of standardization. Item selection and scoring algorithms are standardized, as are testing conditions, instructions, and hardware specifications—so different examinees are treated by the same rules.

This oversimplified capsule of program design in CAT software has made it appear (for the sake of focus) that only item statistics are considered in item choice. Content may be considered, too. If the first few items in a test of elementary arithmetic skill have all been addition items, the program may specify that the next item is chosen not only for an appropriate difficulty parameter or information peak but also for content—it must not

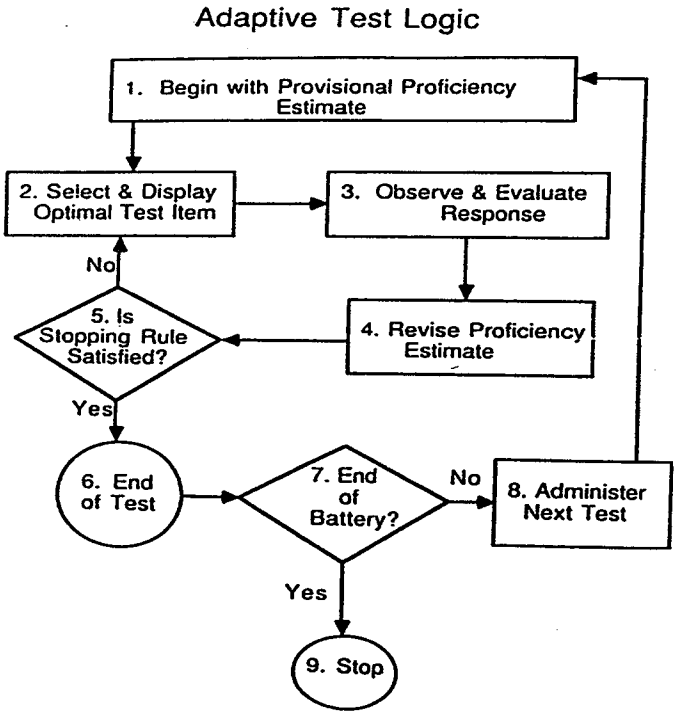

FIG. 11.6. A flowchart describing an adaptive test battery. From Thissen, D., & Mislevy, R. J. (1990). Testing algorithms. In H. Wainer et al. (Eds.), *Computerized adaptive testing: A primer* (pp. 103–135). Hillsdale, NJ: Lawrence Erlbaum Associates. Reprinted with permission.

be another addition item. Moreover, item choices should not result in overuse of some items; some algorithms have allowed some items to be ignored and others to be chosen too often.

Testlets as Items. CAT is long past its infancy, but problems remain. A potentially serious one is the assumption that it does not matter how an item relates to other items. Despite the local independence assumption of IRT, context does matter. Location within a fixed-order test has been shown related to estimated difficulty parameters; so also may information gleaned or highlighted by a different item appearing earlier in the sequence. Despite the important concept of parameter invariance (across samples with differing ability distributions), parameter estimates are unstable enough to be influenced by such context effects (Wainer & Mislevy, 1990).

Wainer and Kiely (1987) suggested the concept of testlets as a test component with more stable characteristics. A *testlet* may consist of perhaps a half-dozen items with homogeneous content. It can be scored; the score is not the simple dichotomy of correct or incorrect responses, but graded response IRT models exist for determining the relevant parameters. The test can be initiated with a testlet of mid-level difficulty, followed by testlets chosen on the basis of score on the first one, and continuing until a stopping rule is satisfied. Testlets can be linear (items arranged in order of increasing difficulty) or hierarchical (branched as in a CAT based on individual items).

An example of a six-item testlet hierarchy is shown in Fig. 11.7. Item numbers reflect the order of difficulty, Item 7 being the hardest. The testlet is hierarchical in that, after the first level, items actually administered are chosen just as they would be in an ordinary CAT. That is, although seven items are in the testlet, the examinee responds to only three of them. Individual item responses are either correct or incorrect; the upper path results from a correct response, the lower one from an incorrect response. The eight different outcomes represent the eight possible patterns of responses to three items actually seen. Each pattern sets an ability estimate for the testlet as a whole.

The first testlet establishes an initial estimate of ability level from which the adaptive program continues from one testlet to another. Because the ability estimates from a testlet are more reliable than those from a single item, the number of testlets required to meet a reasonable stopping target is likely to be fairly small—but the total number of items used will be somewhat larger than in an item-based CAT.

CAT for Personnel Decisions. The discussion of CAT procedures has been brief, partly because of uncertainty about its relevance to personnel assessment and decisions. Adaptive testing can maximize the precision of ability estimation at any point on the ability scale. In personnel

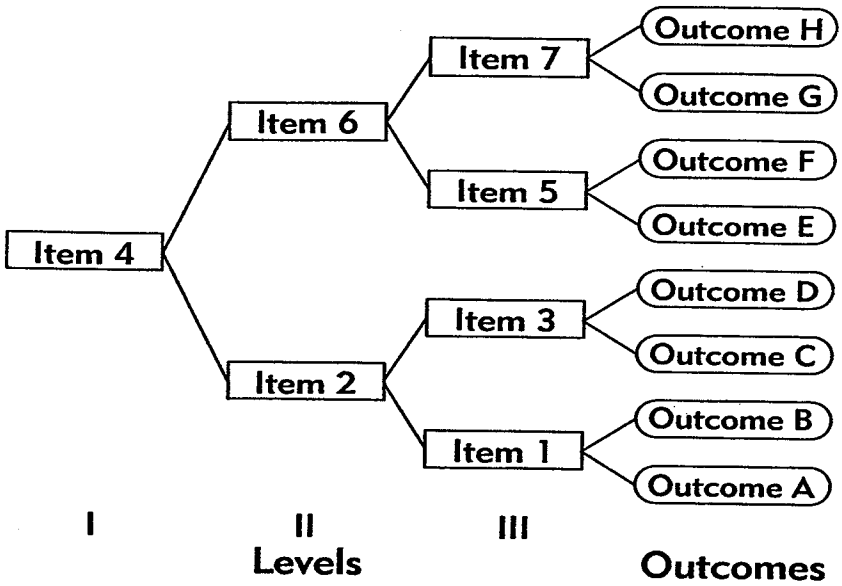

FIG. 11.7. A 3-level, 7-item, hierarchical testlet with eight possible outcomes. From Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185–201. Reprinted with permission.

decisions, however, precision is important mainly at that part of the scale where most decisions are made. If about 20% of those who apply for a job will actually be hired, and most of those offered a job will accept, precise measurement would not be very important below the 75th or above the 90th percentile. With good item parameter estimates, a brief conventional test can be developed that distinguishes well within that narrow region, but not in the low or very high scores where such differentiation amounts to little more than a nice psychometric exercise. One implication of all this is that CAT is probably not useful where cut-score use is likely, as in domain-referenced testing for mastery.

SOME SPECIAL ISSUES

Setting Cut Scores

A cut score effectively dichotomizes a score distribution, loses information, and, if not near the mean, substantially reduces validity. Dichotomization is costly and rarely recommended. Some situations, however, justify and even require a cut score:

1. Civil service jurisdictions commonly give a test to masses of candidates at one time and do not test again for a year or more. Candidates

are listed in an "eligibility list" ordered from those with the highest score to a minimum score. The minimum is a cut score below which examinees may not be listed and no one will be hired. Without a cut score, anyone who sat for the test might eventually be offered a job even if seriously unqualified; the cut score provides one basis (there are others) for deciding when to develop a new exam.

2. Licenses or certification are intended to certify a useful level of knowledge or skill, a degree of competence presumed to protect the public against incompetence. Certification is not limited to governments. Private organizations, including trade associations, may elect to certify the competence or knowledge of sales people, technical advisors, repairers, or others whose work affects customers or the public. A candidate for a job claiming certain expertise may be given performance tests or job knowledge tests to certify that expertise. Licenses and certificates are not awarded on the basis of relative standing in a distribution; in theory, at least, scores are evaluated relative to a prior standard.

3. Hiring may be cyclical. For example, if there is a policy of hiring new graduates from high schools or colleges to work as trainees, most hiring will be done at about graduation time in the spring. Openings may arise at any time through the year. By forecasting the number of openings likely to be needed before the next hiring phase, and with a fairly accurate notion of the score distribution, one can establish a cut score that will provide the necessary number of trainees who can then be assigned to more permanent positions that become available.

4. Assessment may be sequential; an assessment may be scored on a pass–fail dichotomy to decide who gets to the next step. Where many candidates compete for one or a few positions, preliminary screening may be used for all candidates, saving complete assessments (e.g., assessment centers or complex simulations) for the most promising ones. For some jobs, the preliminary assessment may look for intrinsically disqualifying considerations (e.g., poor spelling among proofreader candidates).

Cut scores are too often established merely for convenience. With them, managers getting a candidate's test score need make no judgment more taxing than whether it exceeds the cut point or not—and no HR person need try to explain more valid decision processes to the managers. This bad habit would not be worth mentioning were it not so common, so unnecessary, and so costly in terms of assessment usefulness. I say again,

> A major frustration for me these days is the almost universal and axiomatic use of cutting scores. . . . I'm referring to the kind of cut score above which anyone who comes can be hired and below which no one will be—the kind that changes a continuous score distribution to a dichotomy. A major part

of my frustration is with the reason most often given for setting cut scores: "My managers just can't handle anything more complicated than a pass-or-fail score." *I wish I knew why and when we stopped assuming that decision makers had any brains.* (Guion, 1991a, pp. 14–15)

The *Principles* (Society for Industrial and Organizational Psychology 1987) distinguished between critical score and cutoff score. A *critical score* optimally distinguishes satisfactory (or acceptable, or successful) employees from those who are not. A *cut score* is a decision point, perhaps fluctuating as circumstances change. If applicants abound, it may be higher than a critical score; if they are scarce, it may be lower. A critical score can be used as a cut score, but they are not the same. A related term is *standard*; in educational testing, determining critical or cut scores is called "setting standards."

The Predicted Yield Method. Distributions of candidate qualifications fluctuate from week to week. Availability of openings also vary. The two may not coincide; the best applicants may present themselves when there are no immediate openings. One large company in a small town had such a problem in hiring skilled clerical workers. The best applicants graduated from high school and community colleges in the spring and usually moved away. The solution was to hire good applicants when available, place them in clerical pools, and promote or transfer employees as positions opened up. (That the pool became an excellent training program was an added bonus.)

The plan required fairly accurate prediction of the number of openings likely over the coming year and knowledge of the probable distributions of qualifications. A cut score could then be found to permit hiring enough people at graduation to meet the organization's needs for that year. This kind of cut score is not a costly dichotomization; it is based on a top-down policy. In effect, it is an answer to, "If all these people were available when we wanted them, and if we hired from the top-down as positions opened up, how far down the distribution would we go?"

Thorndike (1949) termed this the *predicted yield policy*. One need not have the limitation of hiring only in the spring to use the predicted yield method, and the time span need not be so long. The need is for reasonably accurate forecasting of positions to be filled and of score distributions. These require good record keeping and research. Number of openings is estimated by knowing of planned retirements and transfers or promotions. Records of past experience with turnover due to sickness, death, or family-related resignations can help. Reasonably accurate forecasts are more likely if informed by research on subsamples; reasons for turnover, for example, may be related to age or sex. Expected organizational changes must also be considered.

Estimating the number of available applicants requires knowledge of economic and employment trends. Local influences should be considered, such as the possible closing of a major business or arrival of a new one. Such factors influence not only an overall number of applicants but the pattern of applicant flow. Test score distributions may be different for different groups of people; they may differ substantially in different local communities. Setting useful cut scores requires realistic knowledge of local distributions, requiring reliable local norms. As time goes by, the original cut score may prove too high or too low to provide the predicted yield—or the predicted number of openings is too high or low—and adjustments may be appropriate.

Regression-Based Methods. Figure 11.8 shows four kinds of relationships. Panel *a* shows a positive, linear regression. Panel *b* shows a positive but nonlinear monotonic regression. In either case, top-down selection is appropriate; a critical score can be based on predicted criterion level.

Panel *c* is a positive monotone up to a point, after which the curve levels off and differences in *X* have no associated differences in *Y*. Above that point, people with different scores should all be considered the same.

Panel *d* (relatively rare) is nonmonotonic. The curve is positive up to a point, after which increases in *X* are associated with criterion *decreases*.

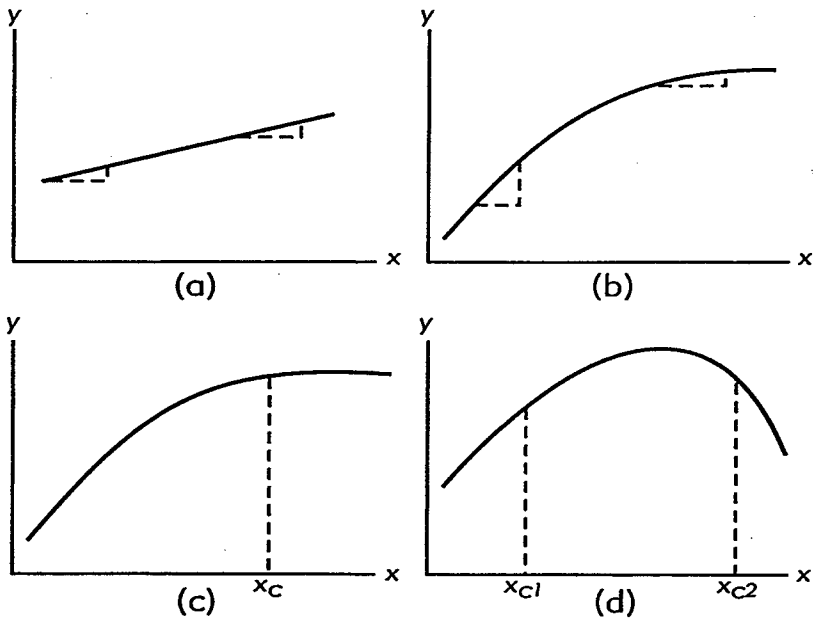

FIG. 11.8. Kinds of relationships of test scores to performance. From Guion, R. M. (1991). Personnel assessment, selection, and placement. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 327–397). Palo Alto, CA: Consulting Psychologists Press.

In such a case, both low and high critical scores might be set to screen out extreme scorers likely to be unsatisfactory. Such patterns seem more likely with personality than with cognitive tests.

In any of these cases, it is possible to base a cut score or two on predicted levels of performance. If a criterion level can be identified that is too low, such as not being able to keep up with a work flow and resulting in lost time for others, the regression equation can be used to identify an associated critical test score or minimum qualification.

Standard Setting by Expert Judgments About Test Items. Very often standards (cut scores) must be set where no regression patterns are known. Methods for setting standards have been developed in educational testing so that performance can be measured against a performance standard rather than normatively (i.e., domain-referenced testing), and more recently, to satisfy state-mandated certification of educational attainment. More information than given here can be found in two major sources: a special issue of the *Journal of Educational Measurement* (Shepard, 1978), and a comprehensive chapter by Jaeger (1989).

Two procedures most often mentioned for setting standards are the Nedelsky and the Angoff procedures. The older Nedelsky procedure applies only to multiple-choice tests. As described by Jaeger (1989), it requires defining, identifying, and sampling a population of judges who are clearly expert in the test subject matter and also understand the kinds of applicants who would be minimally competent in their grasp of the content. *Minimally competent* may be defined in terms of local needs, and may mean minimally acceptable. Each judge must decide which response options a minimally competent person could eliminate as clearly wrong, and then, on the basis of that judgment, compute a minimum pass level for the item—the probability of a correct answer by a minimally competent person. Summing minimum pass levels across items gives the test score the expert expects from a minimally competent person. The mean of these scores across judges is a preliminary cut score.

The Angoff method has the same concerns for defining and drawing from an appropriate population of judges and for conceptualizations of minimally competent persons. It also requires for each item a direct judgment of the probability that the minimally competent person will answer the item correctly. The method is used with multiple-choice, true–false, or short-answer tests, or even work samples scored from checklist items. For each judge, summing these probabilities across all items gives that judge's estimate of the appropriate cut score; an operational cut score could be the mean of the estimates of the several judges.

A modification of the Angoff procedure adjusted judgments by including items previously used (Management Scientists, Inc., 1982). After com-

prehensive training, job experts made judgments about sets of items, mostly new, but including some from previous exams. For each item, they judged the percentages responding correctly among three groups: those whose performance would be unacceptable, acceptable, or better than acceptable. The average of the three judgments was compared to actual difficulty indices of the previously used items, and judgments were adjusted accordingly. Adjusted judgments defined the cut score that would best differentiate the acceptable from the unacceptable. Projected mean and variance were also computed for the new test; projections were so close to actual data that the procedure was accepted as a realistic estimate of an appropriate cut score (*Cuesta v. NY Office of Court Administration, 1987*).

Tests and Controversy

Testing, and personnel assessment generally, is and has been controversial. There are controversies among psychometrically trained experts, among people trained in different test-using disciplines, between psychometric professionals and people outside of these professions, and in society generally. In the face of all the fuss, it is strange that testing remains an important basis for so many kinds of decisions. Few people would want to get rid of various kinds of licensing exams, despite their sometimes serious deficiencies. The cry for educational proficiency exams has been translated into law in many states. Government civil service procedures using merit examination concepts grew out of disenchantment with less objective bases for selection.

In the face of controversy, it is well to remember that tests have compiled a good track record. They have successfully predicted performance on jobs and other kinds of criteria as well. Put together in a battery of tests measuring different things, groups of tests have even better records.

They are good, they are useful, but they are imperfect. Perfection cannot reasonably be expected; too many other things influence criteria for test scores to predict them perfectly. Even so, there is room for improvement. Many things we do well with tests can be done better and with greater understanding. Things we do not do so well with tests provide still greater challenges. The search for new and better ways to measure candidate qualifications, and for new and better definitions of the nature of the qualifying traits, should go forward. However, a lot of bright new ideas, once thought promising, have been tried and have withered. Psychometric history is strewn with the remnants of once grand new ideas. Many tests that were supposed to measure more important constructs than those traditionally measured have gone out of print with only negative findings resulting from their use. Item types once hailed

as panaceas have left the scene in ignominious defeat. Enthusiasm for new ways, commendable as it is, is no substitute for data.

New ideas usually build on old ones. As we approach a new century, there is strong urging for new approaches to measurement and assessment, approaches that do not build on old principles but seek to replace traditional testing with new constructs and methods. Many new ideas are not as different as their enthusiastic proponents assume. Proponents should amass data to show that the expected merits of the new ideas do in fact obtain, that they match or exceed those of the old ones, and that the substitution of the new for the old does not result in losing valued merits of the old without compensating new merits. In short, new ideas in measurement should be sought, articulated, and tried. But we should not allow them to be embraced, adopted, and swallowed whole without competent trial and empirical comparison with the old.

Benefits of Coaching

In most major cities in the United States, organizations exist that purport to teach people how to take tests, especially public sector tests, and get better scores. Such training is called coaching. Does it help?

Professors persist in telling students that they should get ready for the final exam from the very first assignment on, that it represents material learned over a period of time, and that a short session of cramming (make that coaching) cannot compensate for a failure to have had the continuous learning experience. Does this apply to employment tests?

Actually, the question is simplistic, in part because such different things are called coaching. Some coaching teaches people the answers to actual items and perhaps some techniques for answering others like them. Other coaching uses long-term preparatory courses with extended instruction to enhance the abilities or knowledge being measured. Messick (1982) identified three potential aims of coaching: (a) score gain because of test familiarity and subsequent anxiety reduction, (b) gain because of improvement in the skills measured, and (c) gain because of learning test-specific strategies. The latter is probably harmful. The other two may be useful. In reviewing the research, Messick and Jungeblut (1981) found that the amount of coaching time was related to the amount of score improvement. The relationship was not linear; geometric increases in coaching time were accompanied by only equal unit increases in score gain.

To coach or not to coach is an issue not soon to be settled. Many civil service administrators and other employers oppose coaching for their tests, but many entrepreneurs produce new programs. Each new idea for coaching deserves a trial. Nevertheless, my current thinking is that job candidates are not often well-served by coaching programs. I think, ad-

mittedly without adequate evidence, that test familiarization procedures, comparable in logic to realistic job previews, should be provided to try to reduce test anxiety. For the same purpose, I recommend take-home study materials, if test content can be studied. However, such study (cramming) for a few days is not likely to increase actual abilities, or even scores on any but simple tests, and it might increase rather than reduce anxiety. Perhaps the best skill improvement will occur with physical or psychomotor abilities, but the needed training is not a simple matter of a few days with a coach and a set of weights.

Test User Qualifications

Who should buy, handle, and use tests? Are the same qualifications needed for all kinds of tests and inventories? Eyde et al. (1993) developed a book of case studies indexing cases by 86 "elements" of user responsibility and seven factors summarizing those elements (examples are in Fig. 11.9). The booklet includes not only case studies and the full list of 86 elements, but it has much more. Test users should read and follow it.

Translations of Psychometric Instruments

Multinational organizations, and some within a single country, face a special problem in testing people who speak different languages. Mere translation is not the simple matter it would appear. I was once given the example of translating a verbal test item in English into an equivalent French item. The English item depended on the ambiguity of the word *trunk*, meaning either a piece of luggage or the front end of an elephant. No corresponding French word has both meanings. Literal translations, even if possible, may not have the same psychological meaning in two languages; score equivalence is unattainable with literal translation. Translation by "centering" (getting the gist of the meaning) and acceptable back-translation into the original language seems to give equivalent meaning, but that does not assure equivalence in inferences from scores; centering may change psychometric properties dramatically, including constructs measured. Cultural differences can influence scores and their interpretation at least as much as language differences. Cross-cultural testing faces at least three kinds of problems: differences in approaches to tests, problems of test administration, and score equivalence (van de Vijer & Poortinga, 1991).

Two psychometric considerations should govern test translations. First, test item parameters must match in the original and translated versions. Item matching is best done by IRT (Hulin, Drasgow, & Parsons, 1983). Perhaps not every item would be translated to achieve precisely the same

Factor	Elements
1. Comprehensive Assessment Following up testing to get pertinent personal history data to integrate with test scores to enhance accuracy of interpretation.	23. Psychosocial history. 35. Considering patient's state. 37. Teaching research evidence and test limitations. 45. Choice of test to sample relevant behaviors. 77. Follow-up with psychosocial history. 79. Use of tests to generate hypotheses. 82. Proper reporting of clinical observations during testing.
2. Proper Test Use Accepting the responsibility for competent use of the test; exercising appropriate quality control procedures over all aspects of test use.	1. Acceptance of responsibility for competent use of the test. 7. Refraining from helping a favored person earn a good score. 8. Appropriate training and quality control over operations for all users of tests and test results.
3. Psychometric Knowledge Knowing and using correctly basic statistical principles of measurement (e.g., standard error of measurement, reliability, validity).	20. Considering errors of measurement of a test score. 32. Considering the standard error of measurement. 44. Understanding the standard error of measurement.
4. Maintaining Integrity of Test Results Correctly applying psychometric principles to the actual interpretation of test results; understanding the limitations of test scores.	39. Advising administrators about limitations of grade equivalent scores and percentile ranks for specific situations. 49. Making clear that absolute cut-off scores are questionable because they ignore measurement error.
5. Accuracy of Scoring Ensuring that all aspects of test scoring (e.g., recording, checking, correct reading of tables) are performed correctly.	55. Avoiding errors in scoring and recording. 56. Using checks on scoring accuracy. 57. Checking frequently during scoring to catch lapses. 58. Following scoring directions.
6. Appropriate Use of Norms Understanding and using different types of norms correctly, particularly in employment settings.	31. Matching person to job on aptitude validities. 59. Not assuming that a norm for one job applies to a different job.
7. Interpretive Feedback to Clients Providing correct interpretations of test scores to test takers.	71. Willingness to give interpretation and guidance to test takers in counseling situations. 72. Ability to give interpretation and guidance to test takers in counseling situations. 73. Having enough staff to provide counseling.

FIG. 11.9. Seven factors of proper test use with illustrative elements of competent test use. From Eyde, L. D. et al. (1993). *Responsible test use: Case studies for assessing human behavior*. As originally appeared in Eyde, Moreland, Robertson, Primoff, & Most (1988). Washington, DC: American Psychological Association. Copyright by the American Psychological Association. Reprinted with permission.

parameters in a 3-parameter model, but the distributions of item parameters could be kept comparable (Hambleton & Bollwark, 1991). Second, the two versions should be pretty much equally valid measures of the same constructs. Do various antecedent and subsequent correlates behave similarly? Do both versions escape the same contaminating sources of variance? Positive answers say that the tests are measuring the same constructs.

Alternatively, multinational companies can treat operations in each country as independent and develop locally valid assessment procedures. With this option, the entire test development process can take place within

the culture, cultural factors influence construct definition, item writing, instruction development, and all of the developmental research. This option makes sense only if "home country" and local personnel are not competing for the same opportunities, such as promotion to a specified position. Where cross-cultural comparisons are to be made, care must be taken to make the assessments as culturally and psychometrically equivalent as possible.

REFERENCES

American Medical Association. (1977). *Guide to the evaluation of permanent impairment.* Monroe, WI: Author.

Armstrong, A. M. (1993). Cognitive-style differences in testing situations. *Educational Measurement: Issues and Practice, 12*(3), 17–22.

Arnold, J. D., Rauschenberger, J. M., Soubel, W. G., & Guion, R. M. (1982). Validation and utility of a strength test for selecting steelworkers. *Journal of Applied Psychology, 67,* 588–604.

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43,* 695–716.

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. *Personnel Psychology, 27,* 519–533.

Barrett, G. V., & Depinet, R. L. (1991). A reconsideration of testing for competence rather than for intelligence. *American Psychologist, 46,* 1012–1024.

Barton, P. E. (1996). A school-to-work transition system: The role of standards and assessments. In L. B. Resnick & J. G. Wirt (Eds.), *Linking school and work: Roles for standards and assessments* (pp. 125–143). San Francisco: Jossey-Bass.

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology, 76,* 522–532.

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28,* 23–35.

Bennett, R. E. (1993a). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1–27). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bennett, R. E. (1993b). Toward intelligent assessment: An integration of constructed-response testing, artificial intelligence, and model-based measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 99–123). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28,* 77–92.

Bennett, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). *Toward a framework for constructed-response items* (Resch. Rep. RR-90-7). Princeton, NJ: Educational Testing Service.

Bergstrom, B. A., & Gershon, R. C. (1995). Item banking. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 187–204). Lincoln, NE: Buros Institute of Mental Measurement.

Braswell, J., & Kupin, J. (1993). Item formats for assessment in mathematics. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 167–182). Hillsdale, NJ: Lawrence Erlbaum Associates.

Braun, H. I. (1994). Assessing technology in assessment. In E. L. Baker & S. H. O'Neil, Jr. (Eds.), *Technology assessment: Vol. 1, Education and training* (pp. 231–246). Mahwah, NJ: Lawrence Erlbaum Associates.

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1989). *Developing and evaluating a machine-scorable constrained constructed-response item* (Resch. Rep. 89-30. Princeton, NJ: Educational Testing Service.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29,* 253–271.

Cascio, W. F., & Phillips, N. F. (1979). Performance testing: A rose among thorns. *Personnel Psychology, 32,* 751–766.

Chapman, J. C. (1921). *Trade tests.* New York: Holt.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: Wiley.

Cuesta v. State of New York Office of Court Administration, 42 EPD Section 36,949 (SD, NY, 1987).

Curtis, J. R., Gracin, L., & Scott, J. C. (1994, April). *Non-traditional measures for selecting a diverse workforce: A review of four validation studies.* Presented at the meeting of the Society for Industrial and Organizational Psychology, Nashville, TN.

Dalessio, A. T. (1994). Predicting insurance agent turnover using a video-based situational judgment test. *Journal of Business and Psychology, 9,* 23–32.

Dunnette, M. D. (1963). A modified model for test validation and selection research. *Journal of Applied Psychology, 47,* 317–323.

Ebel, R. L. (1972). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Eyde, L. D., Moreland, K. L., Robertson, G. J., Primoff, E. S., & Most, R. B. (1988). *Test user qualifications: A data-based approach to promoting good test use.* Washington, DC: American Psychological Association.

Eyde, L. D., Robertson, G. J., Krug, S. E., Moreland, K. L., Robertson, A. G., Shewan, C. M., Harrison, P. L., Porch, B. E., Hammer, A. L., & Primoff, E. S. (1993). *Responsible test use: Case studies for assessing human behavior.* Washington, DC: American Psychological Association.

Fleishman, E. A. (1988). Some new frontiers in personnel selection research. *Personnel Psychology, 41,* 679–701.

Fleishman, E. A., & Reilly, M. E. (1992). *Handbook of human abilities: Definitions, measurements, and job task requirements.* Palo Alto, CA: Consulting Psychologists Press.

French, J. W. (1965). The relationship of problem-solving styles to the factor composition of tests. *Educational and Psychological Measurement, 25,* 9–28.

Gibson, W, M., & Weiner, J.A. (1996, April). Automated test construction: A novel application of classical test theory. In W. M. Gibson (Chair), *Classical versus IRT methods: Applications in automated test construction.* Symposium at meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18,* 519–521.

Glaser, R. (1994). Criterion-referenced tests: Part I. Origins. *Educational Measurement: Issues and Practice, 13*(4), 9–11.

Glaser, R., & Klaus, D. J. (1962). Proficiency measurement: Assessing human performance. In R. Gagné (Ed.), *Psychological principles in system development* (pp. 421–427). New York: Holt, Rinehart, & Winston.

Goals 2000: Educate America Act (1994, March 31), Public Law 103-227, 108 STAT. 125.

Green, B. F. (1988). Construct validity of computer-based tests. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 77–86). Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, B. F. (1990). System design and operations. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 23–39). Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, B. F. (1991). Guidelines for computer testing. In T. B. Gutkin & S. L. Wise (Eds.), *The computer and the decision-making process* (pp. 245–273). Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, B. F., & Wigdor, A. K. (1991). Measuring job competency. In A. K. Wigdor & B. F. Green, Jr. (Eds.), *Performance assessment for the workplace* (Volume II: Technical issues, pp. 53–74). Washington, DC: National Academy Press.

Guilford, J. P. (1959). *Personality.* New York: McGraw-Hill.

Guion, R. M. (1965). *Personnel testing.* New York: McGraw-Hill.

Guion, R. M. (1979). *Principles of work sample testing: III. Construction and evaluation of work sample tests.* TR-79-A10. Alexandria, VA: United States Army Research Institute for the Behavioral and Social Sciences.

Guion, R. M. (1991a, June). *What I wish I knew about assessment.* Paper presented to the International Personnel Management Association Assessment Council, Chicago, IL.

Guion, R. M. (1991b). Personnel assessment, selection, and placement. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 327–397). Palo Alto, CA: Consulting Psychologists Press.

Guion, R. M. (1996). Evaluation of performance tests for work readiness. In L. R. Resnick & J. G. Wirt (Eds.), *Linking school and work: Roles for standards and assessment* (pp. 267–303). San Francisco: Jossey-Bass.

Hambleton, R. K. (1994). The rise and fall of criterion-referenced measurement? *Educational Measurement: Issues and Practice, 13*(4), 21–26.

Hambleton, R. K., & Bollwark, J. (1991). Adapting tests for use in different cultures: Technical issues and methods. *Bulletin of the International Test Commission, 18*(1,2), 3–32.

Hattrup, K., Schmitt, N., & Landis, R. S. (1992). Equivalence of constructs measured by job-specific and commercially available aptitude tests. *Journal of Applied Psychology, 77,* 298–308.

Hedge, J. W., Teachout, M. S., & Laue, F. J. (1990). *Interview testing as a work sample measure of job proficiency.* AFHRL-TP-89-60. Brooks Air Force Base, TX: Air Force Systems Command.

Hopkins, K. D., & Stanley, J. C. (1981). *Educational and psychological measurement and evaluation* (6th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Horst, P. (1966). *Psychological measurement and prediction.* Belmont, CA: Wadsworth.

Hough, L., & Tippins, N. (1994, April). New designs for selection and placement systems: The *Universal Test Battery.* In N. Schmitt (Chair), *Cutting edge developments in selection.* Symposium at meeting of the Society for Industrial and Organizational Psychology, Nashville, TN.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood, IL: Dow-Jones-Irwin.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York: American Council on Education/Macmillan.

Johnson, D. F., & Mihal, W. L. (1973). Performance of blacks and whites in computerized versus manual testing environments. *American Psychologist, 28,* 694–699.

Koretz, D. (1993). New report on Vermont portfolio project documents challenges. *National Council on Measurement in Education Quarterly Newsletter, 1*(4), 1–2.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5–16.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28,* 563–575.

Link, H. C. (1919). *Employment psychology.* New York: Macmillan.

Linn, R. L. (1994). Criterion-referenced measurement: A valuable perspective clouded by surplus meaning. *Educational Measurement: Issues and Practice, 13*(4), 12–14.

Management Scientists, Inc. (1982). *Development/validation of written examination for uniformed court officer/senior court officer, Office of Court Administration, State of New York* (Vol. III). Philadelphia: MSI.

Martinez, M. E. (1991). A comparison of multiple-choice and constructed figural response items. *Journal of Educational Measurement, 28,* 131–145.

Mazzeo, J., Druesne, B., Raffeld, P. C., Checketts, K. T., & Muhlstein, A. (1991). *Comparability of computer and paper-and-pencil scores for two CLEP general examinations* (College Board Report No. 91-5). New York: College Entrance Examination Board.

McCormick, E. J., & Ilgen, D. R. (1980). *Industrial psychology* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114,* 449–458.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11*(1), 3–9, 20.

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist, 17,* 67–91.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89,* 191–216.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25,* 707–726.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335–366). New York: American Council on Education/Macmillan.

Mislevy, R. J. (1993). A framework for studying differences between multiple-choice and free-response test items. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive testing: Issues in constructed response, performance testing, and portfolio assessment* (pp. 75–106). Hillsdale, NJ: Lawrence Erlbaum Associates.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75,* 640–647.

Motowidlo, S. J., & Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of a situational inventory. *Journal of Occupational and Organizational Psychology, 66,* 337–344.

Olian, J. D. (1984). Genetic screening for employment purposes. *Personnel Psychology, 37,* 423–438.

Osborne, H. F. (1940). Oral trade questions. In W. H. Stead & C. L. Shartle (Eds.), *Occupational counseling techniques: Their development and application* (pp. 30–48). New York: American Book.

Osterlind, S. J. (1989). *Constructing test items.* Boston: Kluwer.

Pike, L. W. (1978, January). *Short-term instruction, testwiseness, and the Scholastic Aptitude Test: A literature review with research recommendations.* Research Bulletin RB 78-2. Princeton, NJ: Educational Testing Service.

Poffenberger, A. T. (1927). *Applied psychology: Its principles and methods.* New York: Appleton.

Popham, W. J. (1974). An approaching peril: Cloud-referenced tests. *Phi Delta Kappan, 56,* 614–615.

Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1994). The instructional consequences of criterion-referenced clarity. *Educational Measurement: Issues and Practice, 13*(4), 15–18, 30.

Reckase, M. D. (1995). Portfolio assessment: A theoretical estimate of score reliability. *Educational Measurement: Issues and Practice, 14*(1), 12–14, 31.

Resnick, L. B., & Wirt, J. G. (Eds.). (1996). *Linking school and work: Roles for standards and assessment.* San Francisco: Jossey-Bass.

Rizzuto, A. P. (1985). *Diazepam and its effects on psychophysiological and behavioral measures of performance.* Doctoral dissertation, Bowling Green State University, Bowling Green, OH.

Robertson, I. T., & Kandola, R. S. (1982). Work sample tests: Validity, adverse impact, and applicant reaction. *Journal of Occupational Psychology, 55,* 171–183.

Schmidt, F. L., Greenthal, A. C., Hunter, J. E., Berner, J. G., & Seaton, F. W. (1977). Job samples vs. paper and pencil trades and technical tests: Adverse impact and examinee attitudes. *Personnel Psychology, 30,* 187–197.

Schmit, M. J., & Ryan, A. M. (1992). Test-taking dispositions: A missing link? *Journal of Applied Psychology, 77,* 629–637.

Schmitt, A. P., & Crocker, L. (1981, April). *Improving examinee performance on multiple-choice tests.* Paper presented at the convention of the American Educational Research Association, Los Angeles.

Schwarz, S. P., McMorris, R. F., & DeMers, L. P. (1991). Reasons for changing answers: An evaluation using personal interviews. *Journal of Educational Measurement, 28,* 163–171.

Shepard, L. (Ed.). (1978). Setting standards [Special issue]. *Journal of Educational Measurement, 15*(4), 237–327.

Shepard, L. (1991). Interview on assessment issues with Lorrie Shepard. *Educational Researcher, 20*(2), 21–23, 27.

Silver, E. M., & Bennett, C. (1987). Modification of the Minnesota Clerical Test to predict performance on video display terminals. *Journal of Applied Psychology, 72,* 153–155.

Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive testing: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45–60). Hillsdale, NJ: Lawrence Erlbaum Associates.

Society for Industrial and Organizational Psychology. (1987). *Principles for the validation and use of personnel selection techniques* (3rd ed.). College Park, MD: Author.

Thissen, D., & Mislevy, R. J. (1990). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 103–135). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement technique.* New York: Wiley.

Thorndike, R. L., & Hagen, E. (1955). *Measurement and evaluation in psychology and education.* New York: Wiley.

Vale, C. D. (1996, April) Generation of equivalent unique conventional test forms. In W. M. Gibson (Chair), *Classical versus IRT methods: Applications in automated test construction.* Symposium at meeting of Society of Industrial and Organizational Psychology, San Diego, CA.

van de Vijer, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277–308). Boston: Kluwer.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–201.

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 65–102). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Thissen, D. (1992). *Combining multiple-choice and constructed response test scores: Toward a Marxist theory of test construction* (Program Statistics Research, Tech. Rep. 92-23). Princeton, NJ: Educational Testing Service.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement, 6,* 1–11.

13

Assessment by Inventories and Interviews

Testing and scaling (including rating) are two basic psychometric procedures; other kinds of assessment procedures are derived from one or both of these approaches. Some of the less constrained constructed response tests are derivatives of both, developed like tests and using rating scales in scoring. Others evolved from the two psychometric foundations and also from forms of assessment that developed outside of the psychometric tradition. Commonly used derivative approaches to assessment, derived both from testing and rating traditions, include inventories and interviews.

INVENTORIES

Inventories are usually self-report measures of interests, motivation, personality, and values. Most of them are developed using test construction principles and, like tests, are scored by summing scores for item responses. Unlike tests, responses are based on opinions, judgments, or attitudes, not on externally verifiable information. Responses may be dichotomous, multiple choice, forced choice, constructed response (as in sentence completion tests), or on rating scales with three or more levels (e.g., agree, uncertain, disagree).

Inventories often have scores for more than one construct; sometimes items for different constructs are mixed throughout a total set. Scoring interpretations may be ipsative rather than normative. The latter is the familiar norm-referenced interpretation expressed as a standard score or

percentile rank. An *ipsative interpretation* compares a person's score on a dimension or construct to that person's scores on the other dimensions (Cattell, 1944). In ability testing, ipsative inference scores would be interpretations of relative strengths and weaknesses; the interpretation in inventories is analogous. It may not be enough in trying to predict turnover, for example, to know that an applicant has a strong need for security; it may be important to know the strength of that need relative to the applicant's other needs such as needs for prestige or self-actualization.

Hughes and Dodd (1961) reported a case where ipsative scoring was valid and normative scoring was not. A stereotyped view of salesmen is that they are highly sociable; normative scoring reflected the stereotype. In their case, however, the salesmen were computer sales people who had to learn a customer's problem and devise a computer system to fit it. Ipsative scoring showed the sociability scale on the Gordon Personal Profile to be negatively related to performance criteria.

Ipsative scoring is relatively rare, partly because of technical problems in using it. Most statistical analyses require operational independence of variables; ipsative scales are not independent. Ipsative and normative scales should not be mixed in regression analysis: in fact, using a set of two or more ipsative variables in a multiple regression analysis is generally unacceptable practice. Moreover, scores to be compared must be on a common scale. If all subscores in an inventory are based on scales developed with the same specifications to produce a common metric, ipsative scoring will work. Scales from different instruments, developed at different times with different people and different specifications, have a common metric only with a common standard score scale, which confounds normative and ipsative measurement. Of course, if this gives valid prediction, practical people will not be upset about it.

Varieties of Inventories

Checklists. Lists of words or phrases can be assembled, and people can be asked to check those that describe them and leave blank those that do not. Items (the words or phrases) usually represent several traits, interspersed to avoid cues to the traits of interest; words can be listed alphabetically, for example; longer phrases can appear in random sequence.

Items might be chosen to fit a theory. Panels of experts may judge whether an item fits a designated trait or not, and a decision rule (e.g., 80% agreement or more) may be set for retaining items. Theory-based checklists are unusual; most are purely empirical. That is, people are classified on an external criterion (e.g., psychiatric judgment, performance level, or staying on the job or not for 2 years), a pool of items is prepared

More theoretically based alternatives exist. Relevant theory might be derived from factor analysis. Biographical scales can be developed to represent certain factors (Russell, 1994), and they will surely be internally consistent. Hough and Paullin (1994) argued, however, that factor scales lose important information and that factor analytic taxonomies are inadequate. They preferred construct-oriented scale construction, beginning with construct definition in the context of a job performance domain and continuing by identifying and defining the trait constructs hypothesized as related to certain aspects of performance. Items should be logically and empirically relevant to the construct as defined and, moreover, should have a kind of face validity, in the sense of obvious relevance to the trait construct; items with keying that go against one's expectations given the trait definition are especially undesirable. Differential item weighting, in their well-supported judgment, is rarely worth the trouble and may cause trouble if the weights are unstable.

Schoenfeldt and Mendoza (1994), however, preferred factor analytic approaches. Factor analysis, especially exploratory analysis, is an empirical approach, but the factors provide rational meaning to their scores. Obviously, factor analytic results depend on the content of the item pool, but factor structures are not purely ad hoc. For example, Table 13.2 lists factors found for customer service occupations; some of these may be job-specific examples of broader constructs, such as the Big Five personality constructs.

The construct-oriented approach advocated by Hough and Paullin (1994) was reflected in the *rainforest empiricism* used by Mael and Hirsch (1993) for one of two biodata forms developed for military academy leadership research. The "rainforest" approach (so-called to isolate it from the pejoratively termed "dustbowl" empiricism) required items with clear relevance to an intended construct, with cumulative empirical data across studies, consistent patterns of item relationships, and multifaceted profiles of criterion performance. The other form, they said, was developed by a quasirational approach. It specified a personality construct, the development of objective personal history items believed relevant to that construct, and items keyed directly to an external personality inventory validly measuring the construct. In short, the quasirational method did not lack empirical data, nor did the empirical method lack rationality. Both forms added incremental predictive validity to existing assessments, although the rainforest approach added more validity with less social desirability effect. Whether called construct-oriented or rainforest empiricism, a combination of data and thought is surely superior to either thoughtless empiricism or naive theorizing.

The "Individual Achievement Record." The development of a biodata form for use in selection for the United States federal government, as described by Gandy, Dye, and MacLane (1994), offers a prototype for an

TABLE 13.2
Factor Scales Derived From a Biodata Form for Occupations With Customer Service Orientation

Dimension	*Number of Items*	*Alpha Reliability*	*Example Item*
1. Sociability	10	0.77	Introduce oneself to strangers
2. Group Membership Participation	10	0.79	Volunteer with service groups
3. Impatience	10	0.76	Upset while waiting
4. Parental Interest	10	0.66	Parents taught hobby
5. Previous Employment	10	0.78	Number of sales jobs
6. Work Ethic	10	0.63	Distracted by family problems at work
7. Male Orientation	11	0.68	Response to competition
8. Work Responsibility	10	0.64	How often late for class in high school
9. Hurry/Accomplishments	11	0.45	Earned major purchase in high school
10. Family Orientation	8	0.36	Assisted with care of family

Note. From Schoenfeldt, L. F., & Mendoza, J. L. (1994). Developing and using factorially derived biographical scales. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 147–169). Palo Alto, CA: CPP Books. Reprinted with permission.

empirical method that is not atheoretical. Development and validation of the form followed these steps:

1. Reviewed information from job analysis of federal nonsupervisory professional and administrative positions and available biodata taxonomies.
2. Established criteria for acceptable biodata items, essentially consistent with Table 13.1 with added concern for use in the public sector.
3. Wrote multiple-choice items with five response options; in most cases, the options represented a quantitative continuum. No preliminary list of constructs guided item development; rather, items reflected "loosely formed hypotheses" that the experiences were related to job performance. Experiences in school, work, and interpersonal areas were included in the pool and believed to tap a variety of constructs. Some items reported factual information; some reported perceptions of the respondent from the perspectives of others.
4. Designated a criterion (supervisory performance ratings) and subjects (entry level professional and administrative people hired over a 4-year period) and collected data.
5. Selected items and developed a scoring key based on double cross validation.

6. Validated scores empirically and analyzed for fairness (using the Cleary method) with data from more than 6,000 employees.
7. Did exploratory and confirmatory factor analyses, identifying four factors among the scored items (also evaluated construct validity by analyzing relationships to reference tests).

Other evaluative studies were done with the completed form, including some designed to promote greater understanding of the factor scores. The project shows the sort of work that can be done with a large sample.

The "Accomplishment Record." Professional people dislike being tested, believing that personnel decisions about them should be based on their records. Lawyers in a federal regulatory agency might also have objected to a test look-alike, for example, a biodata inventory. For them, Hough (1984) developed what she called an *accomplishment record* form and scales.

The critical incident approach to job analysis was used to generate examples of effective and ineffective job behavior; these were sorted by psychologists into dimensions of job performance. An open-ended form was developed for attorneys to use in describing their major accomplishments in each dimension. An example of part of the form is shown in Fig. 13.2.

Responses were scored using BARS. The retranslation procedure was used to assign accomplishment descriptions to the eight dimensions. (Two dimensions seemed confused in the retranslation and were consolidated.) Sixty accomplishments were scaled for each of the resulting seven dimensions by expert judges; descriptions were chosen from those scaled to anchor points on a rating scale, as shown in Fig. 13.3.

The method is time-consuming for researchers, administrators, and examinees, but it is unarguably job related, it does not rely on statistical subtleties, it is a reasonably valid promotion tool, and it seems not to have different effects for men and women or for people of different ethnic groups. In short, it is well worth the time it takes.

INTERVIEWS

Judgments are made during interviews, whether formally recorded as ratings or not, and judgments include assessments, predictions, and decisions. These judgments are often intuitive and haphazard. Assessment may be no more than "sizing up" an interviewee, and prediction may be no more than a vague hunch that the person, as sized up and if hired (retained, promoted, or whatever), will be great, will not be bad, or just would not

USING KNOWLEDGE

Interpreting and synthesizing information to form legal strategies, approaches, lines of argument, etc.; developing new configurations of knowledge, innovative approaches, solutions, strategies, etc., selecting the proper legal theory; using appropriate lines of argument, weighing alternatives and drawing sound conclusions.

Time Period: *1974-75*

General statement of what you accomplished:

I was given the task of transferring our anti-trust investigation of into a coherent set of pleadings presentable to and the Commission for review and approval within the context of the Commission's involvement in shopping centers nationwide.

Description of exactly what you did:

I drafted the complaint and proposed order and wrote the underlying legal memo justifying all charges and proposed remedies. I wrote the memo to the Commission recommending approval of the consent agreement. For the first time, we applied anti-trust principles to this novel factual situation.

Awards or formal recognition:

none

The information verified by: *John* *Compliance*

FIG. 13.2. One dimension of the "Accomplishment Record" inventory and an example of a response. From Hough, L. M. (1984). Development and evaluation of the "accomplishment record" method of selecting and promoting professionals. *Journal of Applied Psychology, 69,* 135–146. Copyright by the American Psychological Association. Reprinted with permission.

work out. Assessments are often secondary to decision; some interviewers want only to reach a decision and then get on with other matters. Herriot (1993) criticized psychometric orientations in interview research; he said that the purpose is to make a decision and that evaluating decisions through statistical prediction is of more interest to academics than to managers. I suppose, with regret, that is true, but good decisions require both competent assessment and explicit (but not necessarily statistical) prediction. Predictions merely implied are rarely articulated or evaluated. My view is that interviews intended for personnel decisions *are* psychometric devices, are based on assessments, and should be evaluated by rules applied to other psychometric devices. Decision making with no concern for quality of assessment and prediction is irresponsible.

Researchers often refer to "the" interview as if all interviews were alike. Just as there are many different tests, there are many different

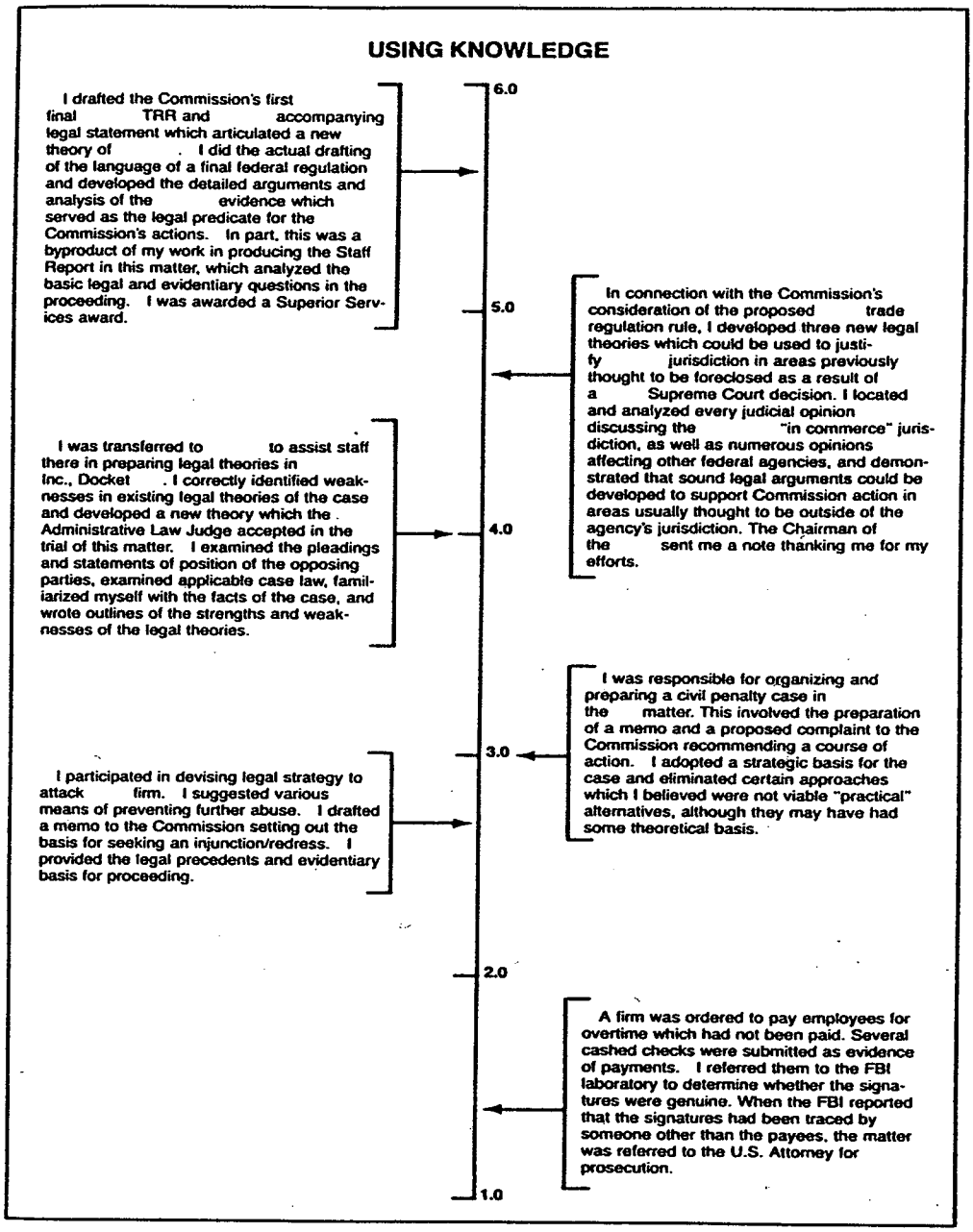

FIG. 13.3. The rating scale for the "Using Knowledge" dimension of the Accomplishment Record Inventory for attorneys. From Hough, L. M. (1984). Development and evaluation of the "accomplishment record" method of selecting and promoting professionals. *Journal of Applied Psychology, 69,* 135–146. Copyright by the American Psychological Association. Reprinted with permission.

interviewers, looking for many different things, and using many different methods. Some are entirely unplanned; others are as tightly structured as any test. Assessment is the avowed purpose of some; it is a hidden purpose in others. Some are short; some are long. Some use one interviewer; others use panels. Some are done by highly skilled interviewers; others are done by people who do not have a clue to useful procedures. Interview content consists partly of the questions or tasks posed and partly of the medium, the individual interviewer. Interviewers are not as standardized as questions; the same questions can be asked in different ways by different interviewers. Stimulus content consists partly of the attitudes interviewers present or the interviewee perceives.

PSYCHOMETRIC EVALUATIONS OF INTERVIEWS

Research Reviews

Interviews have been considered too unreliable to be valid since Hollingworth (1923) reported rank orders assigned to 57 candidates by each of 12 sales managers—with virtually no agreement. A 20-year series of narrative reviews (Mayfield, 1964; Ulrich & Trumbo, 1965; Wagner, 1949; Wright, 1969) consistently identified unreliability as a major problem. Not until Schmitt (1976) was much said about lumping together data from interviewers varying in skill. Early reviewers also tentatively proposed that *structured interviews*, those with pre-planned procedures and sets of questions to be asked, would be better. The idea was later supported in reviews by Arvey and Campion (1982) and Harris (1989).

It would be nice to conclude from the chronology that interviews have improved. I think not. Early research pretty much accepted interviews as they were—haphazard, idiosyncratic, spur-of-the-moment events. The literature was sparse, and reports of even the most ordinary interviewing found their way into it. But so also did the development of interview guides. Hovland and Wonderlic (1939) reported a four-page, 34-question interviewers' guide covering work history, family history, social history, and personal history; McMurry (1947) developed a simpler patterned interview. Others were also developed, and these developments probably influenced researchers without having much influence on the way most interviews were—or still are—conducted: haphazard, idiosyncratic, and spur of the moment. My hunch is that interviews in general are no better but that the literature available for reviewers to survey has improved. If so, we probably know a lot more about assessment by interviewing, and how to make valid, interview-based decisions, than we have communi-

cated to the world at large—where (I suspect) poor interviews remain the rule.

Experimental Studies. Webster (1964) reported a series of experiments at McGill University on variables influencing interviewers' decisions. He and his associates reached several conclusions, among them:

1. Interviewers sharing similar backgrounds develop a stereotype of a good candidate and try to match interviewees and stereotypes.
2. A favorable or unfavorable bias appears early in the interview, and decisions are generally consistent with it. (One finding was that most decisions are actually made within the first 4 minutes of the interview, even if the interview continues well beyond that time.)
3. Interviewers are more impressed by unfavorable information than by favorable information. It is more likely that an early favorable impression will turn into an unfavorable decision than the reverse. Interviewers are "not prepared to take a chance" (Webster, 1964, p. 87).
4. Interviewers seek information to support or refute hypotheses about candidates. When satisfied, they attend to something else.

The McGill studies, well-replicated, were experiments intended to describe the process of reaching a decision, not to evaluate the decision or the reasons for it. It is worth remembering, however, that a decision is itself an assessment; a candidate who gets a favorable decision is deemed ordinally better in some sense than one with an unfavorable decision.

Meta-Analyses. Beginning with Hunter and Hunter (1984), a series of meta-analyses have augmented the narrative reviews and provided explicit generalizations about the validity of (generally) aggregated interviews as predictors of job performance and other criteria. Mean validity coefficients reported in early studies were low but positive; in later analyses, mean coefficients were substantially higher as the literature grew and, perhaps, reported research with better interviews. A reasonable figure is a corrected coefficient (for criterion unreliability and range restriction) of about .36 or .37 (Huffcutt & Arthur, 1994; McDaniel, Whetzel, Schmidt, & Maurer, 1994); Conway, Jako, and Goodman (1996) used upper limits of about .56 for moderately structured interviews and as high as .67 for those that are highly structured—and .34 for poorly structured ones. Interview validity may not be as bad as once believed.

Six possible moderators were studied by Marchese and Muchinsky (1993); the most significant was structure, structured interviews being

more valid. Others were length of interview (longer interviews were less valid) and sex (better validities in pools with mostly female applicants). The authors also correlated moderators with year of publication. More recent articles were more likely to report relatively good validity coefficients, to be based on structured interviews, to be primarily female samples, to be blue collar samples, and to use just one interviewer. The latter may be a function of structure; perhaps the older consensus favoring multiple interviewers was formed when interviews were more likely to be unstructured.

A dramatic difference in mean coefficients (corrected for criterion unreliability and restriction of range) was reported by Wiesner and Cronshaw (1988): .20 for unstructured and .63 for structured interviews. Huffcutt and Arthur (1994) considered structure a probable moderator, dividing it into four levels: (Level 1) wholly unstructured, (Level 2) constraints limited (typically) to topic standardization, (Level 3) prespecifiction of questions (with varying probes allowed), and (Level 4) all applicants asked precisely the same questions without deviations or follow-up questions—essentially an oral test. The four levels of structure, respectively, had mean validities of .20, .35, .56, and .57, all corrected for criterion unreliability and range restriction. The close coefficients at Levels 3 and 4 suggest a point of diminishing returns from structure. McDaniel et al. (1994) reported a mean corrected validity coefficient of .51 for job-related, structured interviews for research (versus administrative) criteria, but removing structure reduced the mean coefficient in that category to .00.

Meta-analytic conclusions evaluate interview validity more favorably than did the narrative reviews. That may be an artifact of the demands of meta-analytic research; a correlation coefficient serving as a data point implies some degree of structure. If validity coefficients for the casual conversations called interviews could be computed, they would probably be lower on average than those with correlation coefficients computed but still called unstructured (McDaniel et al., 1994). My admittedly cynical conclusion: Interviews, if well structured, can be quite valid predictors, but too often are neither structured nor valid.

Varieties of Structured Interviews

It is not easy to define what is meant by structured. Structured versus unstructured is a rhetorical, not a realistic, dichotomy; there are big differences in the degree and the rigidity of structure. In fact, the descriptive term of choice has changed over the years. Wagner (1949) did not call for *structured* interviews; he called for *standardized* interviews. By the time meta-analyses were examining moderators of interview validity, Wagner's term had almost disappeared, although some authors used both terms

interchangeably. They are not synonyms; structure does not necessarily mean standardization. Every time an interviewer decides before an interview what questions will be asked, what judgments will be made, and how they will be recorded, some degree of structure exists; if such structure is developed uniquely for every interview, it is certainly not standardized. It is structured only to fit an individual candidate. It is preparation for the interview, usually done after examining a candidate's credentials—application form, résumé, any letters of recommendation that might be available, and so forth—and noting some concerns worth exploring.

However, the term *structured* more typically refers to interviews tailored to fit a job, not an individual candidate. Structuring in this sense begins with the job description, pay classification, promotion patterns, and related data. From such information, traits relevant to performance may be inferred and appropriate questions identified, to be asked of all candidates. This form of structuring implies at least some standardization.

Different people have different ideas of how interviews should be structured. Four general procedures are described here. The first uses minimal structure, guiding rather than dictating an interviewer's progress through an interview. The second is more tightly structured yet relatively flexible, permitting different candidates to be asked different questions. The other two are more firmly structured, allowing little deviation.

Patterned Interviews. McMurry (1947) developed patterned interviews, a precursor to many lightly structured procedures. It required stating clear, acceptable bases for selection—such as desired traits, background experiences, or training. An interviewer's guide provided kinds of questions that might be asked for each of these, and training was supposed to assure understanding of its questions and the selection standards. Appropriate rating scales were provided for recording summary evaluations. McMurry's rating scales were simple; in a later modification, Maas (1965) used Smith–Kendall scaled expectation ratings for each critical trait.

Behavior Description Interviewing. A more complex modification was called the *Patterned Behavior Description Interview* (Janz, 1989; Janz, Hellervik, & Gilmore, 1986). Janz et al. (1986) gave examples of the interview patterns of questions for 16 jobs. The method is based on the aphorism that the best predictor of future behavior is past behavior; all questions in a pattern ask about past behavior, making it an oral personal history inventory. Question development begins from critical incidents classified into dimensions of behavior. Questions (initial and follow-up questions) are written for each dimension unless that dimension can be assessed better by an alternative to an interview (e.g., tests, biodata,

credentials). The correspondence of question to dimension need not be one-to-one; the same initial question can, with appropriate follow-up probes, provide information for more than one job dimension. For example, a critical incident for an employment test specialist might have been "Developed a valid hands-on performance test to measure problem-solving skills when informed under court order that written tests would not be permitted." The initial question might be, "Tell me about a time when you solved a measurement problem that precluded conventional testing procedures." Follow-up questions might include, "What was unusual about your solution?" and "How did you get your solution accepted by others?" If the job dimensions included creative problem solving and persuasiveness, this question and its probes can tap both. After the interview, the candidate is rated on each job dimension on a simple 5-point graphic rating scale. The sum of the dimension ratings provides a total score.

Situational Interviews. Situational interviews are based on goal-setting theory that states that behavior depends in large part on goals or intentions. Theoretically, if people are asked to say how they would respond to critical situations others have faced on a job, their answers reveal their behavioral intentions. Responses can be systematically scored using a scale anchored by behavioral responses.

Latham (1989) outlined the steps in developing a situational interview:

1. Conduct a job analysis using the critical incident technique. . . .
2. Develop an appraisal instrument such as behavioral observation scales (Latham & Wexley, 1977, 1981) based on the job analysis.
3. Select one or more incidents that formed the basis for the development of performance criteria (e.g., cost consciousness) which constitutes the appraisal instrument.
4. Turn each critical incident into a "what would you do if . . ." question.
5. Develop a scoring guide to facilitate agreement among interviewers on what constitutes a good (5), acceptable (3), or an unacceptable (1) response to each question. If a 2 and 4 anchor can also be developed, do so.
6. Review the questions for comprehensiveness in terms of covering the material identified in the job analysis and summarized on the appraisal instrument.
7. Conduct a pilot study to eliminate questions where applicant/interviewees give the same answers, or where interviewers cannot agree on the scoring.
8. Conduct a criterion-related validity study when feasible to do so. (Latham, 1989, p. 171)

An example of a question and scoring guide is shown in Figure 13.4.

You are in charge of truck drivers in Philadelphia. Your colleague is in charge of truck drivers 800 miles away in Atlanta. Both of you report to the same person. Your salary and bonus are affected 100% by your costs. Your buddy is in desperate need of one of your trucks. If you say no, your costs will remain low and your group will probably win the Golden Flyer award the the quarter. If you say yes, the Atlanta group will probably win this prestigious award because they will make a significant profit for the company. Your boss is preaching costs, costs, costs as well as cooperation with one's peers. Your boss has no control over accounting who are the score keepers. Your boss is highly competitive, he or she rewards winners. You are just as competitive, you are a real winner!

Explain what you would do?

Record answer:

_____

_____

_____

_____

Scoring Guide
(1) I would go for the award. I would explain the circumstances to my buddy and get his or her understanding.
(3) I would get my boss' advice.
(5) I would loan the truck to my buddy. I'd get recognition from my boss and my buddy that I had sacrificed my rear-end for theirs. Then I'd explain the logic to my people.

FIG. 13.4. An example of a question and scoring guide for a situational interview. From Latham, G. P. (1989). The reliability, validity, and practicality of the situational interview. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 169–182). Newbury Park, CA: Sage Publications. Reprinted with permission.

Noteworthy in this sequence is an early focus on performance appraisal, calling for developing criteria first—good advice for any approach. Equally noteworthy is the explicit provision for pilot research. It is noteworthy because people who would never dream of developing written tests without pilot studies often do not hesitate to develop interview guides without them. Building a psychometric device without pilot studies displays unwarranted arrogance—or ignorance of the many things that can go wrong. Among things that can go wrong with this method is anchoring the ends of the 5-point rating scales with examples that do not get used because they are too ridiculous or idealistic. Pilot studies can identify such items.

Like behavior description patterns, situational interviews begin with critical incidents but use them differently. Situational interviews emphasize the future, not the past: "what would you do if . . . ?" rather than "what did you do when . . . ?" Situational interviews usually use panels of two or more interviewers. According to Latham, the typical panel has two managers from the job area and one human resources staff member.

One person reads the questions, but all of them record and evaluate the answers.

Comprehensive Structured Interviews. I have borrowed the term comprehensive structured interview from Harris (1989) to distinguish the specific procedures desribed by Campion, Pursell, and Brown (1988) from the generic term, *structured interview*. Campion et al. (1988) described their procedure as "more highly structured" than most other approaches.

The procedure begins with job analysis to identify KSAs from which interview questions can be developed. Acceptable questions might include those used in behavior description or situational interviews, job knowledge questions, simulations or walk-throughs, and "willingness" questions presenting aspects of realistic job previews. If job requirements differ in importance, the difference is supposed to be reflected by the relative number of questions related to the different ones. Responses are to be rated on 5-point scales, anchored at extremes and the midpoint.

The development of the interview guide does not, at this point, indicate much structure; it offers more freedom than the two preceding methods in studying the job and in the content of the questions. However, the form of the questions is simpler, more like those in a printed test; all candidates are asked precisely the same questions, and no prompting or follow-up questions are permitted (although a question may be repeated if necessary). Moreover, scores of all candidates should be available before the decision is made; this is an explicitly norm-referenced procedure. If feasible, 3-member panels are used; the same panel and the same process is to be used for every candidate. The same panel member is to conduct all interviews and ask all questions; all panel members are to take extensive notes. Questions, answers, and candidates are not to be discussed between interviews, but, after all candidates have been interviewed, large discrepancies in ratings may be discussed and changes made if appropriate. Candidates may not ask questions during the interview, although the procedure calls for a later nonevaluative interview with a personnel representative in which questions are permitted.

Comparison of the Examples. These examples have been presented to show variety, not as prototypes to be matched. All of them have shown reasonable reliabilities and validity coefficients, statistically significant and competitive with other predictors. All have been defended as practical.

There are, of course, unanswered questions. How much structure is necessary? In comparing the four examples, one should keep in mind the diminishing returns of structure as identified by Huffcutt and Arthur (1994). In doing so, however, other questions surface. The most highly

structured interview guides are essentially oral tests with constructed responses. Is test-like standardization an essential feature of interview structure? The same questions could be asked and answered in written form, the responses scored by readers. Would oral and written versions be alike in reliability and validity? Would one form or the other be more susceptible to contaminating sources of variance? Would examinee reaction be the same? We do not know; it is worth investigation. As Hakel (1989) pointed out, we do not know why structured interviews are superior to unstructured interviews, or why they may be about as good as other structured forms of assessment. Again: more research is needed.

Validity

Criterion-Related Validity Coefficients. Interview validity is usually described only with criterion-related validity coefficients; they are apparently higher than formerly supposed. Pooling data across interviewers who differ in individual validity, who make different systematic errors, and whose judgments are not independent, may have seriously underestimated validity coefficients; methods have been offered for unconfounding such data (Dreher, Ash, & Hancock, 1988; Kenny & La Voie, 1985).

Incremental Validity in Prediction. Incremental validity, the increase in variance accounted for when a new predictor is added to those already accounting for some of it, is ordinarily determined by stepwise multiple regression analysis. Tests (and related assessments) are entered first to determine their validity; interview data are entered in the next step to see how much the interview adds. This order of entry arose because tests were more likely to be valid than interviews. As time went by, and structured interviews gave evidence of predictive validity, questions of their incremental validity became important.

Interviews *will* be used, in most organizations, for most jobs. Unless the word gets out about the values of structured interview procedures (Hakel, 1989), many of them will be psychometrically poor. When properly pretested, well-structured interviews are used, however, they usually predict job performance. Are they good enough predictors to add validity?

Maybe not. Walters, Miller, and Ree (1993) developed a structured interview for pilot trainees. It led to seven ratings with, individually, validity coefficients that were modest but comparable to validities for scores on written tests. Equations for written tests with or without the interview ratings did not give significantly different coefficients, perhaps because traits rated by interviewers were also measured by the written tests. Shahani, Dipboye, and Gehrlein (1991) also found no incremental

validity for interview assessments. On the other hand, criteria such as client relations and cooperation were predicted better when interviews were added to the equation reported by Day and Silverman (1989), and interviews provided incremental validity in predicting criteria of leadership, military bearing, and personal discipline (McHenry, Lough, Toquam, Hanson, & Ashworth, 1990).

Sometimes, in my dreams, I reverse the order, asking what incremental validity tests add to interviews. So did Campion, Campion, and Hudson (1994). With a 30-item structured interview and a nine-test battery, they found that interviewer scores added 8% to the criterion variance accounted for by tests; reversing the order in which variables entered the regression equation, tests accounted for only 4% more than interviews alone! Remember, however, that this sort of structured interview is more like an oral test than a conventional interview; maybe it was simply a better test.

Where does such disparate information take us? To the conclusion we already knew: We do not know enough about the incremental validities even of well-structured interviews. Until better accounts of the incremental validities of interviews are available, well-informed decision makers will rely more on tests than on even their own interviewing skills.

Psychometric Validity. Very little attention has been given to the psychometric validities of interviewers' ratings. What inferences, if any, can be validly drawn about interviewees from interviewers' judgments? General answers are unavailable, so no general principles can be offered for improving the meaningfulness of interviews as assessments.

Questions of meaning are questions of constructs and return us to the problem of identifying appropriate constructs for interview assessment. Schmitt (1976) called for research to identify variables best assessed by interviews, but it has not yet happened on any useful scale. Many interviews call for ratings on several dimensions. Sometimes factor analysis of them results in only one factor (e.g., Roth & Campion, 1992). Shahani et al. (1991) reported a factor analysis of interviewer ratings on seven motivation items and five oral communication skill items (the items were developed by committee, apparently without pretesting, which might be the problem). Only one factor accounted for variance among the 12 items. The interviewers could not or perhaps did not distinguish these clearly distinguishable constructs. So what can be said to have been assessed by them?

There are exceptions. Landy (1976) found three factors among nine dimensions (manifest motivation, communication, and personal stability). Rynes and Gerhart (1990) found four factors among 10 dimensions rated.

Nevertheless, they suggested that interviewers were assessing person–job "fit," which, in an understatement, they said is an "elusive construct" (p. 14).

Although interviewer ratings are made in a context different from many other ratings, they are, after all, subject to the problems of other ratings. We will not clearly understand what interviewers can assess until the research enterprise starts to develop theoretical statements of constructs appropriate for interview assessment, train interviewers in their meanings and manifestations, appropriately structure interviews, collect data, and conduct the confirmatory and disconfirmatory research needed to determine whether interviewers' ratings on these constructs lead to valid inferences about them.

Does it matter? Can the decisions be valid even in ignorance of the constructs used in reaching them? Of course. But at this point in the history of employment psychology we should be getting tired of not knowing what we are doing, no matter how carefully we do it.

Content-Oriented Considerations. Interview guides, rating scales, and general structure of interviews are often content-related, relying on job analysis in their development. Lawshe's content validity ratio (CVR; Lawshe, 1975) was computed for items in each of three structured interview guides developed by Carrier, Dalessio, and Brown (1990). One of the guides was for use with experienced applicants, the other two for inexperienced ones. For experienced candidates, the approach worked quite well; the highest CVR items combined to form the best criterion-related validity. Not so for the inexperienced ones. Is content sampling, then, a useful approach to structuring interviews only for experienced people?

I think so. Interview questions and ratings can be informed by the job analysis or derived from it as content samples. The former is like the choice of predictors in a predictive hypothesis and may lead to more appropriate questions for inexperienced applicants. The latter, like work samples and the old oral trade tests, may distinguish truly experienced candidates from those who merely claim the experience. Inexperienced applicants need to be assessed for aptitudes for the work they have yet to learn; aptitude is surely assessed better by tests than by interviewers' ratings.

The Lens Model in Interview Research

Policy capturing and the lens model have shown individual differences in the way interviewers use information to reach overall judgments and in the criterion-related validity of those judgments, and the studies have

shown that treating different interviewers as mere replications of each other (i.e., pooling data across interviewers) is unwise. Some examples:

1. Zedeck, Tziner, and Middlestadt (1983) reported overall validity coefficients of "the" interview (aggregated over 10 interviewers) at barely greater than zero. There were too few cases for individual interviewers to compute corresponding coefficients, but Zedeck et al. showed that individual interviewers had distinctly different decision policies; they wisely concluded that aggregating data (lumping different interviewers together) is inadvisable.

2. In a unique study by Dougherty, Ebert, and Callender (1986), three interviewers audiotaped interviews used in initial screening for entry clerical and technical jobs. Each interviewer saw some applicants and rated them on eight job-related dimensions and on an overall rating scale. All three interviewers rated all applicants from the tapes. Those hired were subsequently rated by their supervisors on ten dimensions, including overall performance. Validity coefficients are shown in Table 13.3. ("Live" judgments are those of the actual interviewer at the time of the interview; all other columns refer to judgments based on the tapes.) Again, aggregated interviewer overall judgments were not significantly correlated with supervisory ratings of overall job performance; neither were ratings from two of the interviewers. The third, however, significantly and substantially predicted all supervisory ratings but one. The study went beyond demonstrating individual differences in interviewer validity; it also showed that interviewers can be trained to use more effective policies.

3. The situation seemed reversed in a study by Kinicki, Lockwood, Hom, and Griffeth (1990). They found significant validity for aggregated data but not for individuals. Again, only a couple of their interviewers had enough cases for appreciable statistical power, and correlations for these disappeared under cross validation.

What conclusion can be drawn? With my predilection for seeing the world as a complex and individualistic environment, I think that policy-capturing research has demonstrated important individual differences in interviewing skill and in the validities of assessments made by individual interviewers; evaluating interviewing by lumping together interviewers who differ in policies and in effectiveness is not useful. Against that predilection, however, must be placed three cautions.

First, the best of these studies used only three interviewers, and very little data from individual interviewers were presented in the others.

Second, where there are several interviewers, clustering techniques show that some interviewers are quite similar in the information they

TABLE 13.3
Validity Coefficients for "Live" Overall Judgments, Mean of Overall Judgments, and Individual Interviewer Judgments

			Interviewer		
Criterion Dimension	*Live*[a] *judgments (n = 57)*	*Mean of*[b] *judgments (n = 57)*	*1 judgment (n = 56)*	*2 judgments (n = 54)*	*3 judgments (n = 56)*
Learning tasks	.10	.17	.09	.07	.24*
Minimal supervision	.05	.32**	.19	.09	.41**
Organizing	.09	.18	.13	−.05	.26*
Judgment	−.05	.24*	.23*	.07	.26*
Job knowledge	−.09	.12	.07	−.11	.23*
Cooperation	−.04	.09	.13	−.01	.08
Productivity	.03	.19	.12	−.05	.32**
Accuracy	.18	.28*	.25*	.19	.27*
Involvement	.06	.28*	.27*	.04	.34**
Overall performance					
Actual	.06	.21	.15	.02	.26*
Predicted[c]			.23*	.19	.26*

[a]Overall judgments made by interviewers in the actual, live interviews; all other columns are correlations based on judgments from the tape recordings. [b]Mean of the judgments based on tapes by the three interviewers. [c]Using judgments predicted from the interviewer's own policy equation.

*$p < .05$; **$p < .01$

Note. Adapted from Dougherty, T. W., Ebert, R. J., & Callender, J. C. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology, 71,* 9–15. Copyright by the American Psychological Association. Reprinted with permission.

consider in making their judgments. Moreover, the similarities can be enhanced by training interviewers to use designated policies.

Finally, meta-analyses have provided no reason to assume serious individual differences, even while making it clear that structure is extremely important. Maybe the importance of individual differences among interviewers is greatly reduced with well-structured interviews.

FACTORS INFLUENCING INTERVIEWER DECISIONS

Interviewer Experience and Habit

Most managers like people with lots of experience, but sometimes we learn things from experience that are not so, including bad habits. Gehrlein, Dipboye, and Shahani (1993) demonstrated that experience is not necessarily helpful to interviewers. Admissions officers (experienced interviewers) interviewed college applicants; other applicants were interviewed by alumni,

faculty, and others termed inexperienced. Validity coefficients of interviewer ratings against GPA were nonsignificant for all of the individual experienced interviewers; surprisingly, inexperienced interviewers did much better. The authors suggested that experience tends to breed confidence even if it is unwarranted. Perhaps the less experienced people compensated for less confidence by planning their interview strategies—in effect, by developing a personal structure for their interviews.

Some interviewers habitually talk too much. Daniels and Otis (1950) found that interviewers generally do most of the talking, sometimes two or three times as much as the interviewees. Moreover, it has been shown that interviewers talk more with applicants they accept (C. W. Anderson, 1960). That finding is hard to interpret. Do interviewers talk more to applicants who show signs of success early in the conversation? Or do they simply feel good about themselves when they talk more, thereby feeling kindly toward the listening applicant?

If the interviewer is seen as an instrument for assessing candidate characteristics through conversation, it seems logical that the interviewer's contributions to the conversation would be relatively brief, encouraging the candidate to speak freely. When the purpose of the interview is to persuade the candidate to accept an offer, perhaps the interviewer should in fact talk more. But in nearly all other purposes, for example, where public relations is to be enhanced, the interviewer is likely to make a better impression on the interviewee by listening than by talking.

Apparently, the amount and kind of talking done by interviewers depends in large part on prior impressions of the candidates. In a decision-making interview, an interviewer often gets prepared by checking out application materials. If this preparation produces a favorable impression, the interviewer is likely to talk more and listen less; there are other first impression effects that bring the validity of interviews into question (Dougherty, Turban, & Callender, 1994).

Experience should lead to skill, not to bad habits. Examining a variety of reviews, Graves (1993) concluded that individual differences in interviewers' skills were likely; she recommended that researchers study individual interviewers to determine what accounts for differences between the effective and less effective ones. She proposed a model of interviewer effectiveness that covers most of the categories of variables described earlier as influencing ratings. She also gave 19 propositions for a research agenda. From some of these, I offer four principles that, admitting the need for research, seem supportable enough to be followed (the language is mine, and she might not approve):

1. Interviewer effectiveness depends on the richness of the interviewer's job-related cognitive structure, and this richness depends

on the experience of the interviewer. In addition to thorough training and supervised experience, interviewers should accumulate a wealth of other experiences to enrich their understanding of the jobs to be filled.

2. Interviewers should be bright, intelligent, analytical people.
3. Interviewers should have a clear, job relevant prototype of an ideal applicant for a job to be filled (more on this later). Interviewers might be encouraged to deviate modestly from the prototype, but not to make extreme deviations. That is, they should not be too rigid about it, but should look for people who fit the prototype reasonably well.
4. The interview should be structured and include only topics of clear job relevance.

Nonverbal Cues

Much has been written about nonverbal communication, especially as used by interviewees to make a good impression on interviewers (N. R. Anderson, 1991). Interviewers should know that they can be unduly influenced by such behavior, but many of them base judgments on it, anyway. Experienced (not necessarily good) interviewers have told me that they rely on a variety of nonverbal candidate behaviors: leaning back after making a statement (reason for distrust), firm handshake (strong character), catching one's breath (sign of lying), clean clothes (sign of neat work habits). They have not, however, given me validity evidence. Neither has research. At present, at least, interviewer reliance on interviewee nonverbal behavior must be treated as a potential source of error.

Stereotypes, Prototypes, and Biases

In the Netherlands, Van Vianen and Willemsen (1992) asked 307 employees in university scientific and technical jobs to check adjectives describing attributes identified in job advertisements. One group of subjects filled out the checklist from the point of view of evaluating a future colleague. The other half responded to the items as generally associated with men, with women, or with both genders. An item was classed as "masculine" or "feminine" or "sex neutral" by rather stringent criteria. The final, abbreviated list was dubbed the "Sex Stereotype Attribute List." Its scoring key, applied to future colleague evaluations, showed gender stereotypes for ideal applicants for various jobs, and interviewer judgments were consistent with them.

The notion of an ideal applicant need not be stereotypic. Prototypes of ideal candidates can be developed by deliberation, perhaps from job descriptions or with the help of supervisors and senior employees.

> How do different interviewers develop and use prototypes of desired candidates? . . . I distinguish between a stereotype (which develops willy-nilly, is widely accepted, and seems implicitly to apply to all members of a group) and a prototype, by which I mean something like a car designer's prototype, a carefully and systematically developed ideal to be achieved; for selection, the prototype should be defined by a set of attributes that not only describe the desired candidates but distinguish them from those less desired. . . . I suspect that work on the idea of a prototype as a planned ideal will be more fruitful than work on more or less generally accepted stereotypes of what is." (Guion, 1987, p. 202)

"Similar-to-me" is a bias. "Similar-to-ideal candidate" seems a useful match to an ideal prototype; if the prototype is valid, matching it should imply valid assessment as well. After interviews, Dalessio and Imada (1984) asked each interviewer in a five-member panel to complete a descriptive rating form including seven college majors, 10 personality traits, 11 interests, and six preferences. Three weeks after all interviews were completed, the interviewers filled out the same form describing (a) an ideal applicant and (b) themselves. Actual interview decisions were more closely related to the ideal applicant match than to the self-applicant match.

Interviewers' biases potentially include demographic variables like sex, race, ethnicity, or age, although research generally reports little or nonsignificant differences in interviewers' ratings of men and women, or of different ethnic groups. However, a more general "similar-to-me" bias could inflate tendencies toward bias. In one study, racial similarity effects were stronger in conventional than in structured interviews, although mixed-race panels of interviewers avoided the effect (Lin, Dobbins, & Farh, 1992); similarity effects were not found for age. Another study of panels of interviewers showed a similar racial effect, giving higher ratings to candidates of the same racial identity as the majority of the panel (Prewett-Livingston, Feild, Veres, & Lewis, 1996).

Interviewee Characteristics

Obviously, characteristics of the person interviewed should influence decisions; they include the characteristics sought. Two special cases, however, merit concern as potential sources of error.

Memory. Interviews generally consist of questions requiring the interviewee to respond with a remembered event, state, or behavior. Personal recall may not be accurate (Pearson, Ross, & Dawes, 1992). People may have

have implicit theories of their own personalities that emphasize stability (e.g., This is how I think now, so I must have thought similarly then). Other people, or the same people for other questions, have implicit theories that lead them to exaggerate changes that have occurred. That is, people make the implicit assumption that behaviors match attitudes. If a person recalls behavior (e.g., leaving a job) associated with an attitude, and if the attitude has changed, the response may describe behavior more in line with the present attitude than with the earlier reality.

Impression Management. Candidates try to make good impressions, and some are better at it than others. *Impression management* is the attempt to influence the impression made on others. There are surely individual differences in self-presentation skills, but there is little information about kinds of job performance these skills may predict or the kinds of assessments they may contaminate (Fletcher, 1990). Interview research needs to study the effect of impression management. Does behavior successfully creating the desired impressions with one interviewer work equally well with another? Can interviewers learn to detect the deceptions the term "impression management" implies? If so, can they successfully ignore it in making job-relevant assessments or decisions. In a widely cited article, Kinicki et al. (1990) found that two factors described interviewer ratings on six dimensions. One they labeled "interview impression," the other was called "relevant qualifications." The terms are adequately descriptive; only the relevant qualifications factor validly predicted independent job performance ratings.

In General

A large body of research on interviewing has, in my opinion, given too little practical information about how to structure an interview, how to conduct it, and how to use it as an assessment device. I think I know from the research that (a) interviews can be valid, (b) for validity they require structuring and standardization, (c) that structure, like many other things, can be carried too far, (d) that without carefully planned structure (and maybe even with it) interviewers talk too much, and (e) that the interviews made routinely in nearly every organization could be vastly improved if interviewers were aware of and used these conclusions. There is more to be learned and applied.

REFERENCES

Albright, L. E., Glennon, J. R., & Siegert, P. A. (1963). Measuring achievement motivation at the time of employment. *Journal of Industrial Psychology, 1*, 59–65.

Alliger, G. M., Lilienfeld, S. O., & Mitchell, K. E. (1996). The susceptibility of overt and covert integrity tests to coaching and faking. *Psychological Science, 7*, 32–39.